POLISH ACADEMY OF SCIENCES
COMMITTEE OF TRANSPORT

ARCHIVES OF TRANSPORT

QUARTERLY

ARCHIWUM TRANSPORTU

ISSN 0866-9546

volume 43
issue 3
Warsaw 2017

Detailed instructions for authors and contacts to editors can be found on journal's webpage: www.archivesoftransport.com

The Archive of Transport is indexed by the Polish Ministry of Science and Higher Education.

All articles are peer-reviewed by two external reviewers. Reviewers list is published once each year in the last issue of the journal and also on journal's webpage.

The journal is supported by the Polish Academy of Sciences using funds of the Polish Ministry of Science and Higher Education.

PAN
POLSKA AKADEMIA NAUK

Printed by:
Centrum Poligrafii Sp. z o.o., ul. Łopuszańska 53, 02-232 Warszawa.

Circulation 165 copies.

ARCHIVES OF TRANSPORT

Associate Editors (Subject Editors)

Editorial Committee Contact:

ARCHIVES OF TRANSPORT
Faculty of Transport, Warsaw University of Technology
Koszykowa 75, 00-662 Warsaw
Poland
E-mail: archivessecretary@wt.pw.edu.pl
Phone: +48 22 234-58-55

Contents

ARCHIVES OF TRANSPORT www.czasopisma.pan.pl PAN POLSKA AKADEMIA NAUK www.journals.pan.pl ***ISSN (print):*** *0866-9546*
Volume 43, Issue 3, 2017 ***e-ISSN (online):*** *2300-8830*
DOI: *10.5604/01.3001.0010.4224*

MODELLING THE PUBLIC TRANSPORT CAPACITY CONSTRAINTS' IMPACT ON PASSENGER PATH CHOICES IN TRANSIT ASSIGNMENT MODELS

Arkadiusz Drabicki[1], Rafał Kucharski[2], Andrzej Szarata[3]

[1,2,3] Cracow University of Technology, Faculty of Civil Engineering, Department of Transportation Systems, Cracow, Poland

[1]e-mail: adrabicki@pk.edu.pl
[2]e-mail: rkucharski@pk.edu.pl
[3]e-mail: aszarata@pk.edu.pl

Abstract: *The objective of this paper is to discuss the replication of passenger congestion (overcrowding) effects on output path choices in public transport assignment models. Based on a comprehensive literature review, the impact of passenger overcrowding effects was summarised in 3 main categories: the inclusion of physical capacity constraints (limits); the feedback effect between transport demand and supply performance; and the feedback effect on travel cost (discomfort penalty). Further on, sample case studies are presented, which prove that the inclusion of capacity constraints might significantly influence the assignment output and overall results in public transport projects' assessment – yet most state-of-the-practice assignment models would either miss or neglect these overcrowding-induced phenomena.*

In a classical 4-step demand model, their impact on passengers' travelling strategies is often limited to path (route) choice stage, while in reality they also have far-reaching implications for modal choices, temporal choices and long-term demand adaptation processes. This notion has been investigated in numerous research works, leading to different assignment approaches to account for impact of public transport capacity constraints – a simplified, implicit approach (implemented in macroscopic-based models, e.g. PTV VISUM), and a more complex, explicit approach (incorporated in mesoscopic-based models, e.g. BusMezzo). In the simulation part of this paper, sample tests performed on a small-scale network aim to provide a general comparison between these two approaches and arising differences in the assignment output. The implicit approach reveals some differences in assignment output once network capacity constraints are accounted for – though in a simplified manner, and producing somewhat ambiguous output (e.g. in higher congestion scenarios). The explicit approach provides a more accurate representation of overcrowding-induced phenomena - especially the evolving demand-supply interactions in the event of arising congestion in the public transport network. Further studies should involve tests on a city-scale, multimodal transport model, as well as empirical model validation, in order to fully assess the effectiveness of these distinct assignment approaches.

Highlights:

- *The paper discusses the inclusion of overcrowding effects on path choices in public transport assignment models*
- *These can be grouped into 3 main categories: physical constraints, demand-supply feedback and path discomfort cost*
- *Sample case studies show that their inclusion may substantially affect the assignment output*
- *Two general methods of modelling capacity constraints are: the implicit and explicit approach*
- *An illustrative example shows that both approaches produce different output with the explicit one being more specific and adequate*

Key words: *public transport assignment, passenger congestion, overcrowding, crowding discomfort, path choice, public transport capacity.*

1. Introduction

Path choice (or route choice) process comprises a crucial step within every single transport assignment model (Fig. 1). The path (route) choice algorithm is most commonly described by means of the probabilistic, discrete choice model and the random utility theory (Cascetta, 2001). The bottom line is that the probability of choosing a given O-D path is related to the cost-utility formula, which reflects the relative (dis)utility associated with travelling along that path, among all the alternative O-D paths (routes). The path cost formula comprises the following trip components: perceived travel times (i.e. in-vehicle, waiting, walking times), monetary costs (fares), transfer penalties and temporal utilities of earlier (or later) O-D connections – which are described in relative (weighted) terms, reflecting the user perceptions of disutility associated with particular trip stages (e.g. increased disutility associated with waiting and walking times). This path cost evaluation algorithm forms a key component within the classical 4-stage assignment modelling framework, where it is applicable at the modal split stage – i.e., used to evaluate the choice probability between the public and private transport modes (Szarata, 2014) – and eventually at the trip assignment stage – i.e., used to compute the choice probability of feasible network paths (routes, lines etc.).

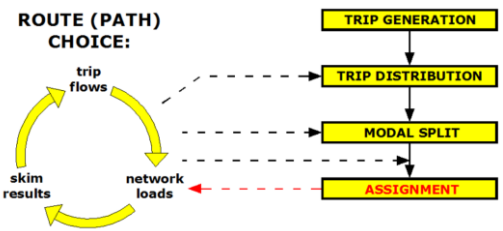


Fig. 1. Path (route) choice process in the 4-stage assignment model (source: Hartl, 2013)

In a summary, this means that the passenger choices in public transport networks are principally a function of journey times and service frequencies – i.e. main factors which are recurrently deemed most important according to the passenger surveys (Rudnicki, 1999). However, a major factor which is either missing or not properly exploited in most state-of-the-practice assignment models, concerns the inclusion of line (service) capacity, as well as the associated (dis)comfort aspects – which might actually also have a notable effect on passengers' output choices, especially when considering various public transport modes with distinct capacity rates - i.e. mass transit (underground, urban rail) vs. feeder (light rail) systems, or conventional (bus, tram) vs. unconventional modes (monorail) (Drabicki et al., 2016).

The implications of network capacity constraints (limits) are more investigated in case of private transport assignment (Żochowska, 2014), where they are typically included in form of the volume-delay function (VDF). The VDF function describes the effects of increasing travel times as a result of rising traffic flows (volumes) – i.e. a non-linear travel time penalty which increases sharply once traffic volume approaches the saturation flow rate (i.e. the assumed road capacity limit). However, whereas the VDF functions are commonly available and widely applied to replicate the capacity limits in modern-day private transport (PrT) assignment models (Branston, 1976), the incorporation of capacity constraints effects in public transport (PuT) assignment models remains – to the best of our knowledge – much less examined and advanced, usually limited to individual case studies and modelling developments; in practical approach, the implications of capacity constraints on passenger path choices are often neglected in state-of-the-practice modelling algorithms.

The objective of this paper is to contribute to the on-going research discussion on replicating the overcrowding effects in public transport assignment models. The literature review part of this paper will outline the main aspects of their impact on passenger path choices (which should be accounted for in simulation models) and present sample results from practical transportation studies. Further on, the simulation works on a small-scale network will reveal the arising differences in assignment output between the two common modelling approaches to public transport capacity constraints. Our aim is that the observations and conclusions from this study would illustrate the possibility of reproducing the overcrowding effects in these two main modelling algorithms, provide indications for their application on bigger-scale transport models – and together with a summary of the state-of-the-art in public transport congestion modelling, it would also point out fields for future improvement works.

The remainder of this paper is organised in a following way: section 2 focuses on literature review regarding the incorporation of public transport network capacity constraints and their impact on output passenger path choices. Section 3 highlights the importance of proper appraisal of public transport capacity constraints effects in assignment models, presenting results from sample case studies, where capacity constraints were taken into account and led to substantially different project assessment indicators. Section 4 presents two distinct modelling algorithms of public transport networks, where the influence of capacity constraints can be described in 2 various approaches. These are followed by practical simulations on sample networks in section 5, where both modelling approaches lead to different network performance, and consequently – distinct simulation output. Finally, section 6 provides the summary and conclusions for further research works, and indications for future applications on bigger, city-scale public transport assignment models.

2. Literature review

Substantial amount of research works in recent years has been devoted to the notion of public transport congestion, or more precisely – the passenger overcrowding: i.e., the way it affects the passengers' travelling strategies, user preferences, implications for the transport system performance, the issues of service optimisation etc. - with an ultimate goal of the inclusion of these (often mutually dependent) effects in assignment models. However, though these state-of-the-art assignment models aim to replicate the impact of passenger overcrowding on path choice decision models in a most plausible way possible, they usually include only some of the overcrowding effects. Consequently, they often do not yield completely realistic results and are likely to underestimate the arising phenomena of passenger overcrowding.

In a general overview, the effects of passenger overcrowding on output path choice decisions – which should be accounted for in a model observing the public transport capacity constraints - can be summarised into three main categories, as listed below and elaborated in subsequent sections:

- physical capacity limits (constraints),
- feedback effect on service performance,
- feedback effect on passenger (dis)comfort.

2.1. Passenger congestion effects in assignment model – impact of physical capacity limits

The first category of public transport congestion effects concerns the direct impact of **physical capacity constraints** – i.e., the maximum permissible flow volume of passengers which can be carried by the components of public transport network within a specified time period. The major factor determining the physical capacity limits is the passenger load capacity of public transport vehicles (the max. no. of passengers able to "get on-board") and the arising queuing phenomena at stops or platforms – while (Gentile and Noekel, 2015) also suggest that in some cases the finite capacity limits of stops (platforms) themselves – i.e. the space limitations – are also of relevant importance. In recent literature works, the impact of physical capacity limits in public transport assignment has been typically modelled in 2 following ways: by means of the (so-called) implicit or explicit approaches.

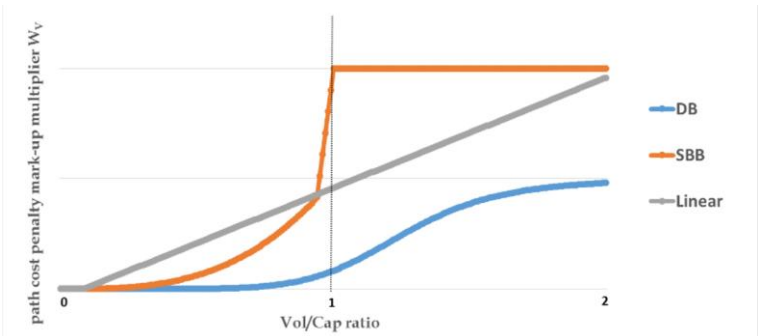


Fig. 2. Crowding (mark-up) cost discomfort functions, available in the implicit approach (based on (PTV VISUM Manual, 2016))

The **implicit approach** to capacity constraints follows the VDF-based method used as a default in most private transport assignment models to represent the congestion effects (described earlier). The passenger flow capacity of network links is not strictly bounded by a fixed limit value, but is instead defined with a non-decreasing, volume-dependent link cost function (Fig. 2). Typically, the function imposes an additional cost penalty above a certain threshold (e.g. the assumed seat capacity), which reflects the rising crowding discomfort. Further on, as passenger volume tends towards the capacity limit, the so-called *crush capacity*, the increase in cost penalty becomes non-linear and very sharp. Once the passenger flow exceeds the nominal line capacity (i.e. calculated with respect to the crush capacity of operating vehicles), travellers are not explicitly prohibited from using the service (vehicle run), but travel cost should have now risen so severely, that any additional passengers should be "discouraged" from boarding it – i.e., an *implicit* capacity limit is imposed upon that particular service (vehicle run). Analogous to the capacity-constrained traffic assignment model, the assignment is calculated in an iterative procedure: in each consecutive simulation run, the output demand flows (i.e. travellers' choices) depend concurrently on network parameters (i.e. travelling conditions) calculated in preceding simulation run. The assignment procedure gradually converges towards a stable solution, and the final output (passenger flows, line loads and travel costs) is obtained once an equilibrium state is achieved.

Typically, the path cost penalty in implicit approach (as e.g. in PTV VISUM model) is described either as a linear function of the volume-to-capacity ratio, or utilises a more nuanced, non-linear correlation as e.g. assumed by the DB and SBB functions (fig. 2) – with the latter solution being perhaps more appropriate, as it allows to account for the non-uniform increase rate in crowding discomfort. The two non-linear crowding cost functions used in the PTV VISUM model are analogous to the approaches used in rail demand modelling in the German railway system – the DB function (*Deutsche Bahn*), and the Swiss railway system - the SBB function (*Schweizerische Bundesbahnen*). In these two functions the path cost penalty due to passenger overcrowding is in general exponentially correlated with the rising volume-to-capacity ratio, with an upper bound limit of the crowding cost penalty rate – beyond which it converges towards a fixed penalty rate (typically, this would occur once the *crush capacity* limit has been reached).

This forms a simplified method of representing the effects of passenger congestion in public transport assignment, which is usually applied within macroscopic models and available in common transport modelling tools (e.g. (PTV VISUM Manual, 2016)). As shown below on sample case studies, this assignment method enables to replicate some effects on passenger overcrowding on route choices and modal shifts, yet it comprises a rather simplified approach (e.g. by imposing a uniform cost both for travellers on-board and those waiting at the stops), missing the important, evolving congestion phenomena in public transport system.

The **explicit approach** to capacity constraints comprises a more specific (and thus more reliable) representation of public transport supply and its interactions with travel demand. Though it has not been applied yet on a wider scale – being developed mostly in individual algorithms (e.g. the BusMezzo algorithm (Cats, 2011)) and case study applications - its implementation has been hitherto possible in mesoscopic and microscopic assignment models. A more detailed modelling framework implies that the travel demand is represented by individual agents (passengers) progressing through the network, whereas travel supply is represented by individual vehicles (runs) defined with strict capacity limits, corresponding to the crush capacity values. Travellers arriving at the platform (stop) board the incoming vehicle runs according to their residual (available) capacity. If boarding volume exceeds the residual vehicle capacity, the remaining passengers are explicitly denied the boarding and have to wait for next vehicle departures – thus, important queuing phenomena arise at the platforms (stops). The queuing discipline at stops can be commonly reproduced in a number of ways, notably including the following two (Gentile and Noekel, 2016):

- the FIFO principle: "first in, first out" – an organised queuing process, consisting of the undersaturation queue (those who will board the nearest vehicle run) and oversaturation queue (those delayed and "forced" to wait yet for later vehicle runs),
- the so-called *mingling process*: no priority rules are in place, and passengers joining the residual

queue have roughly the same boarding probability as others waiting at the platform.

The resultant fail-to-board probability, which becomes significant as passenger congestion rises, has wider implications on the ensuing passenger path choices (Nuzzolo et al., 2012). Travellers who had to skip previous vehicle runs perceive additional disutility due to the boarding failure – i.e. the arising waiting cost is perceived as relatively more burdensome. Consequently, they may take a rerouting decision and consider other, less attractive O-D travel routes (paths).

2.2. Passenger congestion effects in assignment model – feedback on service performance

The second type of passenger overcrowding effects, resultant from the inclusion of public transport capacity limits, concerns the feedback interaction between the transport supply (service regularity and dwell times) and transport demand (passengers' decisions and resultant volume flows) performance. A principal reason underlying this interaction is that the dwell-time of a public transport service trip is an increasing function of boarding and alighting passenger volumes. In a summary, the feedback effect demonstrates itself in the following manner: changes in passenger flows cause fluctuations in dwelling times at stops, which will conversely induce variations in service operating times and headway deviations. In turn, as vehicle arrivals (and departures) become irregular, passenger demand is now unevenly distributed among the individual runs – and further on, the feedback effect is amplified. This impedes the service regularity and reliability which is undesirable both for passengers (increasing travel times and crowding levels) as well as for operators (uneven utilisation of service supply).

This important phenomenon, "reinforced" by the arising passenger overcrowding, can only be replicated if the modelling framework allows to describe the impact of demand flows on vehicle dwell times – which is in practice often neglected especially in macroscopic assignment algorithms. The boarding and alighting processes are strictly related to passenger flow vector, which depends on vehicle exchange capacity, and the assumed "dwelling routine" (i.e. separate or mixed doors for boarding and alighting). Based on a wide range of literature sources (summarised by (Tirachini et al., 2013), (Gentile and Noekel, 2016)), it can be concluded that there is a roughly linear correlation between the dwell times and number of alighting (boarding) passengers – the values fall usually within the range of 2-4 secs/pass, though these are likely to increase even further (up to 6 secs/pass and beyond) in overcrowded conditions. (Gentile and Noekel, 2016) provide a detailed mathematical framework for describing the impact of dwelling flows on mean and, crucially, variance values of dwell times and service headways – i.e. the main "trigger" behind this feedback interaction process. Importantly, these time-dependent service variations may initially occur at individual stops or line sections, but will likely become amplified and propagate further downstream in the network.

The feedback loop between transport demand and transport supply performance is probably best manifested in a well-known phenomenon, which occurs in public transport networks during congested conditions – i.e. the so-called **bus bunching** effect (Fig. 3); its other denominations mentioned in literature sources are: bus platooning, clumping, pairing, the banana bus, the Bangkok effect (Moreira-Matias et al., 2012).

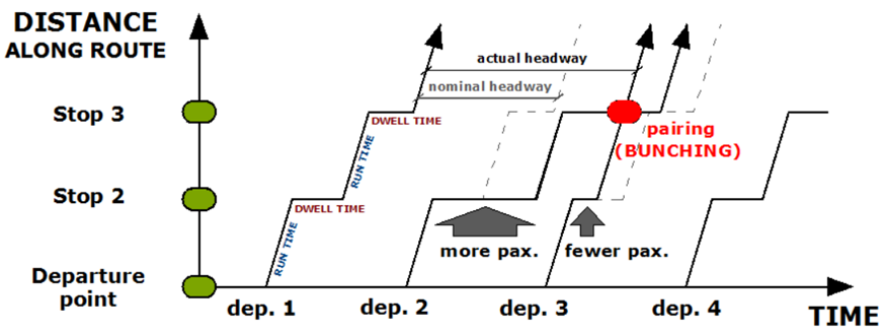


Fig. 3. Bus bunching effect, plotted on the space-time diagram (source: Attanucci, 2010)

The bus bunching effect can be explained intuitively on a space-time diagram (Attanucci, 2010), under the simple assumption of constant (Poisson-distribution based) passenger arrival process at stops, as follows: a certain vehicle run which arrives later than scheduled at the stop has to pick up a higher than average number of waiting passengers. Dwelling time takes longer than expected and once ready to depart, the vehicle is now delayed even further (relative to its nominal timetable). The same pattern will hold at the next downstream stop, where overcrowding conditions will likely become worse, the service delay will rise further, and so on. In contrast, the next (following) vehicle run has less waiting passengers to pick up, dwells shorter at the stop, and as a result, will run ahead of schedule. The relative headway between these 2 consecutive runs will likely decrease as they progress downstream in the network, and the second vehicle run may eventually catch up with the vehicle ahead of it – the stage where vehicle runs become "fully bunched" or paired together, at which the relative headway drops down to zero. The bunching phenomenon leads to substantial impairment in public transport service regularity, since journey times become longer, the waiting times are higher (due to uneven vehicle spacing), and recurrently – the average crowding levels increase due to uneven passenger loads' distribution among the individual vehicle runs.

One of the objectives of research works was to describe the main factors and critical conditions which induce the bus bunching effect. A common conclusion is that passenger demand (volume) has profound impact on service regularity, or more specifically, the resultant loading factor, defined as the ratio of pass. arrival rate (at stop) to pass. loading rate (on-board). (Newell and Potts, 1964) developed (possibly) a first mathematical framework for bunching effect, where they define this correlation by means of a critical bus bunching parameter. It describes transition from stable conditions to a self-reinforcing bunching phenomenon state, at which – if sustained over a longer time period - the buses will fall out of schedule even further. More advanced approaches emphasise the importance of passenger arrival pattern, which need not be always uniformly distributed in time. For example, (Fonzone et al., 2015) demonstrate on the proposed algorithm that various possible arrival patterns would require different critical conditions to trigger the bunching effect (which could then develop in a substantially distinct degree). (Gentile and Noekel, 2015) propose a bus bunching coefficient variable, defined as a function of service headway between 2 consecutive vehicle runs. The coefficient can be used as a basic measure of arising bus bunching effect in the network, being calculated as the ratio of actual service headway (i.e. one resulting from fluctuations in dwell times) to the nominal scheduled headway – the higher the headway deviation rate, the bigger the on-going bunching effect. An analogous formula can be used to estimate the bunching coefficient at a downstream stop, resulting from current upstream service conditions and passenger dwelling flows.

Additionally, research sources mention that the bus bunching effect is not only related to the demand-supply interactions at the stops, but may also be induced (or amplified) by other factors, such as general traffic characteristics, route design, road conditions etc. Numerous analytical models have been developed which allow to demonstrate their impact upon the output service regularity ((e.g. (Bąk, 2010), (Horbachov et al., 2015)) – however, in this paper we will focus primarily on the influence of passenger congestion and the consequent bunching phenomena.

2.3. Passenger congestion effects in assignment model – feedback on passenger discomfort

The third major type of public transport congestion effects concerns the arising discomfort cost and its implications for passengers' travelling choices. Evidence from passenger surveys seems to reinforce the fact that crowding (dis)comfort is among the major factors relevant to users' travel experience – e.g. results from Transport for London's regular monitoring of customer satisfaction (Barry, 2015) indicate that travel comfort and crowding are rated as the (third and fourth) most important issues, right after the journey time and personal safety. Although journey times still form the baseline and most decisive factor in path choice process, the travel discomfort may also contribute its own mark-up "penalty" upon the travel cost. Overcrowding affects the travellers' comfort perception who become more reluctant if their public transport services are routinely congested. (Tirachini et al., 2013) mention a wide range of psychological, sensorial and social factors attributed to the overcrowding effects, such as: risk perception of personal safety, anxiety and

stress, possible ill-health, propensity to arrive late at work, possible loss of productive time.

Commonly, the **crowding discomfort factor** is included as an additional (mark-up) travel time multiplier in the general path cost formula. The relative (perceived) value of travel time components increases as rising passenger numbers (flows) produce a crowding externality (cost), relative to travelling in uncrowded conditions. The discomfort penalty is described as a non-linear, VDF-based function of volume-to-capacity ratio of a given travel alternative, which increases more sharply as crowding conditions deteriorate – the generalised crowding mark-up factor formula (Gentile and Noekel, 2015) is based on the same VDF function as used in the implicit approach to modelling the capacity constraints (described earlier).

A recurring question in literature sources is how to measure precisely the on-board crowding levels, with two basic approaches considered (Tirachini et al., 2013):

- discomfort cost as a function of *load factor* (percentage volume-to-capacity ratio): a simplified measure which can be related to the vehicle seat capacity, or in macroscopic approach – roughly to the generated line capacity – yet it says very little about the actual on-board crowding conditions themselves, which will vary depending on (among others) the vehicle interior arrangement; studies estimate that as such the crowding cost is "activated" from load factors between 60 – 90% onwards (Tirachini et al., 2013), (van Oort et al., 2015),
- discomfort cost as a function of *density of standees* (per square metre): perhaps a more relevant measure, since crowding discomfort becomes much more acute once passenger load surpasses the vehicle seat capacity, and the estimated available space per passenger provides a better picture of the degree of crowding "suffered" by standing travellers; here, the crowding mark-up penalty ranges between 1.0 – 1.6 (for those seated) and 1.5 – 2.4 (for those standing), and applies already if density of standees rises from zero (pax. per sq. m) (Whelan and Crockett, 2009).

The exact values of crowding discomfort factor differ among literature sources, being dependent on the methodology used, local context and user preferences, as well as individual public transport modes and trip characteristics (Tirachini et al., 2013). Literature review shows that crowding discomfort values are likely to be higher in case of rail systems and increase with trip length and duration. The majority of studies which aimed to provide an estimate of crowding discomfort costs on passengers' choices focused mainly on long-distance urban trips (i.e. between the suburbs and city centre) made with suburban or metro railways (Tirachini et al., 2013), (Kroes et al., 2013), (Whelan and Crockett, 2009), as well as intercity rail trips (Lieberherr and Pritscher, 2012). A meta-study commissioned for the UK Department of Transport (Whelan and Crockett, 2009) provides a comprehensive valuation of overcrowding costs and the willingness-to-pay estimate for trips made in the British Rail system – which are often used as a guideline in transport practice (Fig. 4).

Table 4.2: Crowding Value of Time Multipliers

Load Factor	Sit	Stand	pass/m^2	Sit	Stand
80	1.00	1.50	0	1.00	1.53
100	1.08	1.50	1.0	1.11	1.62
120	1.23	1.67	2.0	1.21	1.70
140	1.38	1.85	3.0	1.32	1.79
160	1.53	2.02	4.0	1.42	1.87
180	1.68	2.20	5.0	1.53	1.96
200	1.83	2.37	6.0	1.63	2.04

Please note: The Load Factor and pass/m^2 (passengers per square meter) estimates vary by rolling stock type. The rows in this table are therefore **do not** match across different crowding metrics.

Fig. 4. Time cost multiplier factor due to crowding - acc. to the British Rail WTP meta-study (source: Whelan and Crockett, 2009)

The purpose of such modelling framework is to reflect the crowding discomfort impact on a certain share of travellers who would adjust their travel patterns, so as to avoid the worst overcrowding circumstances - and utilise other O-D travel alternatives. This should replicate the long-term adaptation process in travelling strategies, as repeated experience of overcrowding will impact the 3 important aspects:

- path (route) choice: travellers will be less likely to use notoriously overcrowded services and would seek other, perhaps less attractive, public transport connections; in the modelling approach, this could imply a demand shift towards services with higher spare capacity (e.g. mass transit systems), or less-popular public transport connections (e.g. a trade-off in longer journey times combined with less-crowded travel conditions),
- mode choice: as a consequence of routine overcrowding, public transport service would lose on their relative attractiveness, and travellers would likely revert to using private cars; for short-range trips, possibly an increase in walking (or cycling) trips could be observed,
- departure time choice: perhaps the most significant impact of passenger congestion on travel choices - in the day-to-day adaptation process travellers will seek to avoid the time periods of peak congestion, and would utilise the same O-D travel route but at less-popular travel times; the departure-time updating process would imply a higher passenger volume migration especially towards earlier departure runs.

This adaptability phenomenon of passengers' path choice strategies in response to crowding discomfort can be incorporated in the modelling framework by means of a conventional, iterative user equilibrium approach (in a simplified manner), or more reliably – by employing a day-to-day learning mechanism (Nuzzolo et al., 2012). In the latter case, travellers consider on day t the anticipated attribute values of path cost components, which are a weighted average of experienced and anticipated attribute values on day t-1 – thus, the path choice model is recurrently updated based on users' expectations and their prior experience:

The extent to which the overcrowding experience impacts the passengers' choices will differ profoundly, depending on the trip purpose (motivation). Literature sources (Tirachini et al., 2013) and empirical surveys alike (London Assembly Report, 2009a) confirm that crowding discomfort is of limited significance for commuter (obligatory) journeys but might have substantial implications for leisure (non-obligatory) journeys. In former case, the necessity to arrive at destination on-time means that commuter travellers still assign much higher (relative) weight to travel times than on-board conditions, or as given by a cited London commuter (London Assembly Report, 2009b): "*You just have to use the Tube. There's just no choice, there is no option. Well, there is an option: just don't go to work but that's not really an option!*". The same report examines the ways in which commuters adapt to the frequently experienced travel conditions: around 66% of London rail commuters adjusted their departure times (e.g. chose earlier connections), and for one of the rail services ca. 20% of travellers would travel in the opposite direction first, just to have a higher chance of getting a seat at an upstream station. On the other hand, crowding seems to have much more suppressive impact on the non-obligatory trip motivations. The majority of leisure travellers do avoid travelling on London Underground during rush hours, and 25% of them change the time of travel during the day due to anticipated crowding. Additionally, sociodemographic factors themselves might be relevant as well: (Kim et al., 2009) indicate that specific user groups expose different "sensitivity" rates to crowding – e.g. elderly people are likely to sacrifice the extra travel time in favour of more comfortable trip conditions.

3. Appraisal of public transport capacity – sample case studies

Incorporation of passenger overcrowding effects on public transport system has been shown in a number of (both academic and practical) case studies to influence the overall projected network usage, performance results and assessment indicators – yielding distinct results when compared to the analysis "insensitive" to overcrowding phenomena. (Batarce et al., 2015) point out interestingly that (passenger) congestion in public transport plays an analogous role to (traffic) congestion in private transport (Fig. 5): investment in public transport systems increase both their transportation capacity and relative attractiveness, which spurs passenger demand growth. However, in longer run this induces

increase in travelling discomfort due to arising crowding, and the (finite) capacity of public transport supply itself may eventually become outstripped by the ever increasing passenger demand. In the end, this implies that further improvements in public transport systems are necessary - the ramifications of this feedback correlation may only be captured if public transport capacity constraints are taken into account; neglecting it would produce erroneous results in terms of public transport system effectiveness and capability.

Both implicit and explicit approaches to modelling the capacity constraints have been utilised in sample case studies to demonstrate the arising differences in public transport assignment output between congested vs. uncongested cases. In a recent case study for The Hague city (van Oort et al., 2015), an implicit, VDF-based approach revealed differences in passenger flows' distribution between the proposed tram line and the existing bus route along the same transport corridor (Fig. 6). A two-tier crowding mark-up penalty was assigned to the path cost formula, which reflected first an increasing discomfort penalty due to rising on-board crowding (within the range of 1.0 – 1.7), and after reaching the assumed *crush capacity* it surged rapidly up to the constant value of 10.0. The method revealed a higher patronage rate of the tramway system – the passenger gains could be attributed both to its higher nominal capacity limit as well as better on-board comfort level, when compared to the existing bus system. Importantly, a reduction in service frequency need not necessarily imply a decline in passenger numbers, as envisaged by uncongested model. Inclusion of another principal factor – i.e. increasing service capacity (provided by tramway system) – mitigated these losses and even projected a slightly higher demand flow along the proposed tram line.

Another case study in the city of Stockholm (Cats et al., 2015) utilised a more specific, explicit approach with individually modelled vehicles and travellers (agents) to assess the projected performance of a new metro line proposed along an existing, busy bus corridor.

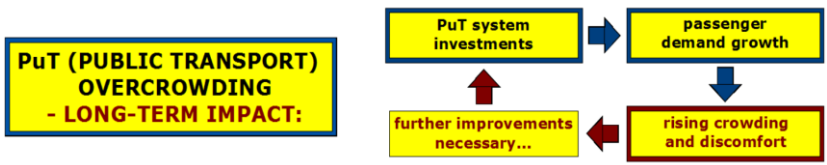


Fig. 5. Public transport (PuT) congestion (overcrowding) - a long-term feedback impact which may not be captured with conventional assignment models (source: Batarce et al., 2015)

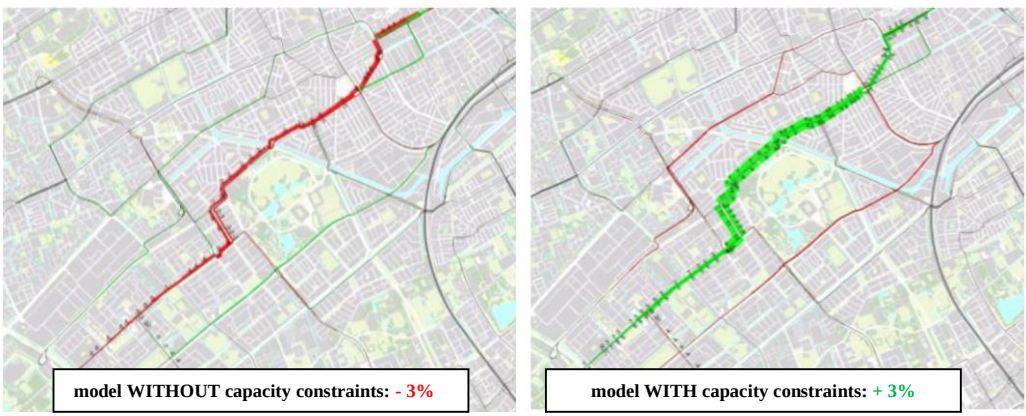


Fig. 6. Sample results of including the capacity constraints' and comfort effects in the implicit approach – estimated relative effect on daily ridership after conversion of bus line 25 to tram line in The Hague city (source: (van Oort et al., 2015))

In the existing scenario, when the busy corridor is served by 200 buses per hour (per direction), ca. 40% of buses are routinely overcrowded, and multiple denial-of-boarding events can be observed. In that case, the explicit approach captures the deteriorations in service quality and travel times, in the form of the bus bunching effect caused by (and correlated with) excessive demand flows – i.e. the very principal ramifications of the mutual demand-supply interactions. In contrast, a new metro line with much higher capacity would attract ca. 60% of bus users, and despite smaller service frequency it would still be less overcrowded and much more resilient to service disruptions. The absolute decrease in in-vehicle and waiting times is ca. 15%, but when weighted in relative (perceived) terms, the project would bring ca. 65% extra benefits, attributable to higher system capacity and reduced discomfort travel cost. In such case, a cost-benefit analysis based on uncongested static model would potentially miss a major share of gains coming from public transport system improvements.

A study for the Swiss railway system (Lieberherr and Pritscher, 2012) developed an implicit, VDF-based capacity restraint model, the so-called SBB crowding function (described in more detail in subsequent chapter), which has been now incorporated in the macroscopic PTV VISUM software. Application in pilot projects showed that the capacity-restraint assignment reduced the overestimation (overload) rate of railway system usage (measured in seat-km) by 30% - though the assignment model would still yield somewhat overestimated passenger flows, the obtained results would be more plausible. Additionally, the SBB crowding function revealed extra shifts from intercity to regional train services during overcrowded peak hours – a minor share of travellers (ca. 3% of total O-D flow) would switch towards slower but less-crowded trains. Furthermore, researchers reckon that a more far-reaching distinction between "seated" and "standing" crowding penalty itself might influence the assignment output. (Leurent, 2009) demonstrate that the predicted passenger load in Paris metro system is reduced by ca. 30%, when a congested model additionally distinguishes between the seated and standing crowding disutility.

Additionally, researchers (Small, 1999) indicate that the benefits of improving the public transport system capacity may be not only quickly diminished (i.e. "eaten-up") by passenger influx from alternative routes (modes), but furthermore – they might be actually partially (or even totally) undone by the phenomenon of latent (induced) demand (Szarata, 2013). The city of London provides a good example in terms of that narrative, illustrating how massive investment programmes in transportation systems can barely keep up with the ever growing demand pressure. A multi-billion improvement programme currently underway across the London Tube (underground rail) system is projected to increase the system capacity by approx. 30%, but analysis prepared for the busiest Tube line, the Northern Line (Fig. 7), shows already that by the time the works have been finished - the crowding levels will be even worse than before, virtually along each single section of the line (Transport for London, 2013). A flagship Crossrail project (ca. £17bn of total cost) is supposed to contribute 10% to the total urban transport network capacity – a substantial nominal gain in the city of 8m inhabitants - and become a core part of the public transport system. Though it is widely expected to relieve the existing Tube network, transport planners predict that once opened in 2018 the Crossrail "will be immediately full up with people" (Drabicki., 2015), and argue that a second Crossrail line is badly "needed" to counteract the anticipated passenger congestion. Numerous similar case studies can be found elsewhere in biggest urban metropolitan areas across the world, in case of which the public transport systems are particularly likely to become prone to massive passenger congestion and induced demand pressure.

As mentioned earlier, impact of overcrowding on long-term passenger path choices also concerns the modal choices and departure time choices. (Tirachini et al., 2013) use stated-preference passenger survey data, and propose a range of MNL models to estimate demand choice models arising from inclusion of crowding discomfort in travel cost formula (Fig. 8). This Sydney-based study emphasises that models insensitive to crowding discomfort are likely to underestimate the value of in-vehicle travel times savings and overestimate the demand (model) share for high congestion levels (and vice versa for low congestion levels). An important observation is that for suburban railway trips, the inclusion of repeated overcrowding experience should produce a demand shift towards

private transport, with crowding "sensitivity" rate increasing as a function of trip duration: an uncongested model would yield a constant modal share of a sample rail line at ca. 5%, whereas for a congested model the modal share would range between 4 – 6% (travel time of 15 minutes) or ca. 3 – 8% (travel time of 40 minutes). In terms of departure time updating process, (Nuzzolo et al., 2012) incorporate a day-to-day learning mechanism in public transport assignment model, so as to emphasise the long-term implications of crowding experience. The proposed framework shows that approx. 65% of commuters shift towards other (earlier or later) vehicle runs to mitigate the risk of on-board congestion, leaving on average 5 minutes earlier at the origin – a "spillback" effect can be observed in temporal demand distribution pattern: individual vehicle run loads might now substantially differ from their initial values once a congestion-induced adjustment takes place in passengers' choice process.

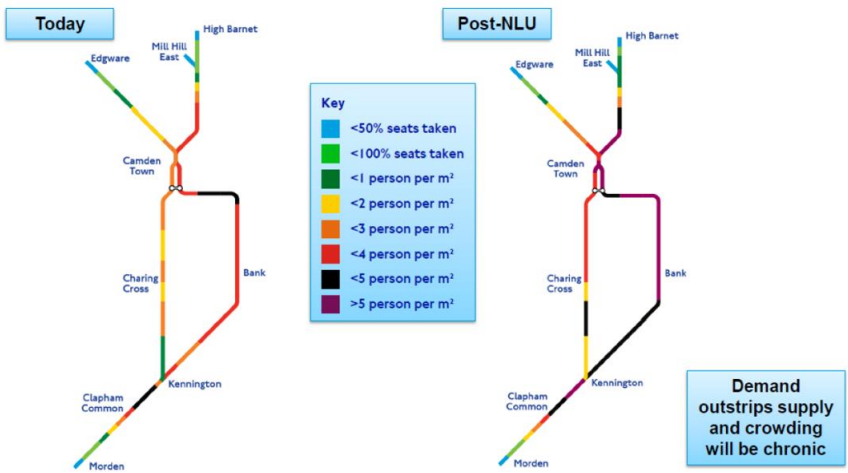


Fig. 7. Sample results for the London Underground case study: despite massive investment programme (NLU), capacity increases on the Northern Line will be quickly absorbed by induced passenger demand growth (source: Transport for London, 2013)

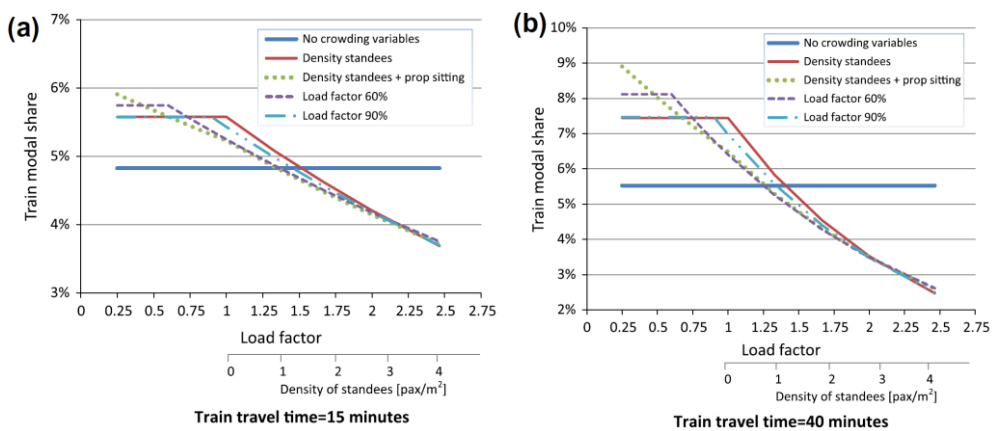


Fig. 8. Sample results obtained with different crowding cost functions – correlation between the overcrowding (discomfort) impact and the modal share of commuter rail system (source: Tirachini et al., 2013)

4. Assignment algorithms

In the practical part of this study, two public transport assignment algorithms will be tested on a sample transport network, to observe how they replicate the effects of passenger overcrowding on the evolving transport system performance and travelling experience (mainly in terms of journey times and service loads). Each algorithm utilises a distinct approach to modelling the capacity constraints of public transport systems, and assumes a different modelling aggregation level both on demand and supply sides, i.e.:

- **implicit approach**: timetable-based (i.e. schedule-based), macroscopic assignment model – as implemented in the commonly-used PTV VISUM software,
- **explicit approach**: simulation-based (i.e. agent-based), mesoscopic assignment model – as incorporated in the currently developed BusMezzo software.

4.1. Implicit capacity constraints' algorithm

The timetable-based assignment model operates on a macroscopic level, reproducing travel demand in form of aggregated link flows (within a certain time period). The path choice model is a one-off process triggered at the origin, when traveller chooses a complete O-D path (route), based on its (predetermined) utility value – and follows that single path all the way to his (her) destination. The baseline path utility formula is a sum of weighted (perceived) travel time components (in-vehicle, waiting, walking times), transfer penalties, and the temporal utility of that O-D connection. Additionally, once a capacity restraint model is introduced, a crowding mark-up penalty *(1 +Av)* is assigned to the total path utility. The crowding penalty is recalculated in an iterative process, based on the volume-to-capacity ratio of each link segment (importantly – not individual line segments), until a certain convergence (equilibrium) threshold is attained – i.e. a fixed-point problem solution after which a final path utility (impedance) rate is evaluated. The algorithm utilises a VDF-based procedure analogous to the private transport congested assignment model, with 3 crowding impedance functions available. For the purposes of this study, the SBB (Swiss Railway) function was assumed as it should allow us to replicate the two-tier effect of rising network overcrowding upon the path utility (impedance): for low volume-to-capacity rates, the effects of rising passenger discomfort (crowding mark-up penalty within range of 1.0 – 1.7), and a step-wise jump in path impedance (constant crowding mark-up penalty of 10.0) once passenger volume exceeds the assumed *crush capacity* (Fig. 9). This algorithm should reproduce, in a simplified – i.e. *implicit* – approach, the effects of capacity constraints on output passenger path choices: reductions in excessive (overestimated) passenger volumes and increasing attractiveness of less-crowded routes – though without considering the more specific, congestion-associated phenomena, particularly at the stops.

4.2. Explicit capacity constraints' algorithm

The simulation-based assignment model assumes a more disaggregate representation both of transport demand – individual agents (travellers), and transport supply – individual vehicles (trips) operating within the network. Here, the path utility is recurrently updated at each journey stage, when traveller may reconsider his (her) path (route) choice towards the destination – i.e. at each instance a boarding, alighting or connection decision process is triggered.

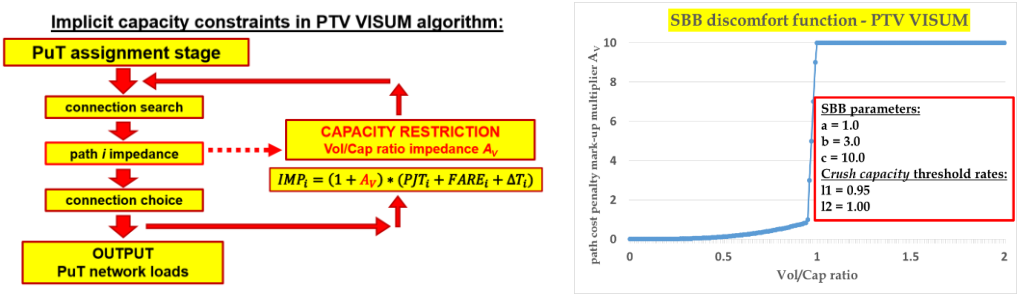


Fig. 9. Implicit capacity constraints' algorithm assumptions in simulations

Likewise, the path utility formula comprises the same set of travel time components plus transfer penalties, except for temporal utility of connection which was not yet included in the algorithm. Since the modelling algorithm operates on a more detailed, mesoscopic level, resultant passenger flows are an aggregate output of all the individual actions (path choice decisions) taken by agents (travellers) progressing through the network. Service supply is modelled as individual trips (runs) served by public transport vehicles, which are described with their distinguished properties - vehicle type, vehicle dynamics, and importantly – specified maximum passenger load capacity. Network performance is reproduced in a more stochastic manner – the actual travel times depend on real-time system conditions, and a dwell-time function is introduced to describe the direct impact of dwelling (i.e. boarding and alighting) flows onto dwell times – in our case, we will assume a linear dwell-time function of 2 secs/pass. The utilised modelling framework did not incorporate yet the impact of crowding discomfort upon the path cost-utility formula; nonetheless, it would allow us to observe the actual transport network performance and its implications for passengers' travelling experience once network capacity constraints are modelled in an *explicit* approach – i.e. with strict denial-of-boarding and arising queuing phenomena occurring at stops if passenger flows exceed the system capacity, and the very important demand-supply interactions (Fig. 10).

5. Results – implicit and explicit approaches

Simulations presented below were performed on a sample public transport network, i.e. the extended version ("SF ENet" (Fonzone and Schmoecker, 2014)) of the classical Spiess-Florian network (Spiess and Florian, 1989). The extended SF ENet layout is assumed on a network topology formulated by (Fonzone, Schmoecker 2015) and comprises a system of 7 bus stops (A to G) and 5 unidirectional bus lines (L1 to L5), situated along 2 parallel O-D routes (Fig. 11). A single origin-destination pair is assigned to the network. Travellers are allowed to transfer between bus lines at stops, and additionally a two-way, 3-minute walking connection is provided between the intermediate stops C and F. Vehicle runs are dispatched from origin stops at fixed intervals (headways), and line run times between consecutive stops remain constant. The crush capacity rate of each bus vehicle is assumed as 100 pax.; in explicit approach, a dwell-time function is introduced with a linear rate of 2 secs/pass.

To analyse the incorporation of passenger overcrowding effects in the sample network, 2 distinct modelling approaches were included, i.e. the implicit approach (PTV VISUM) and the explicit approach (BusMezzo) to modelling the capacity constraints.

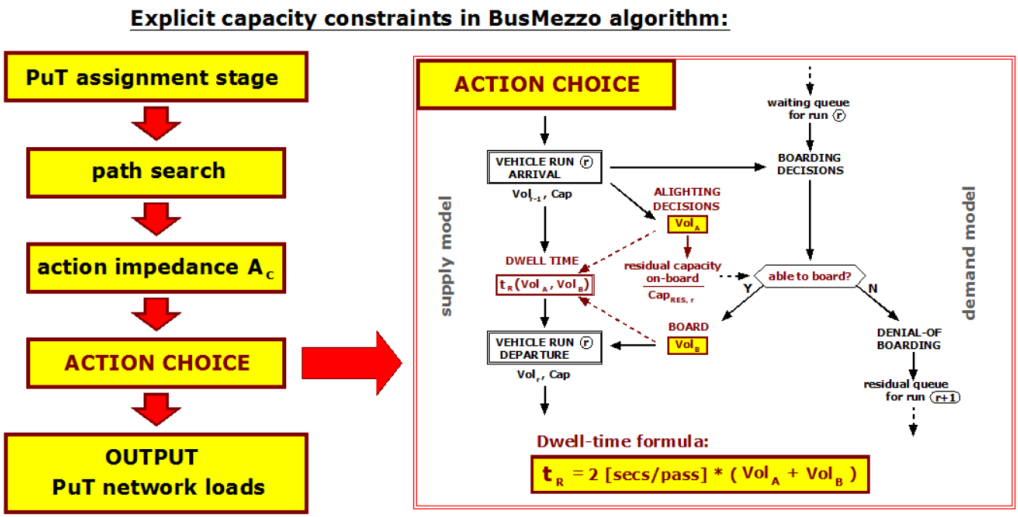


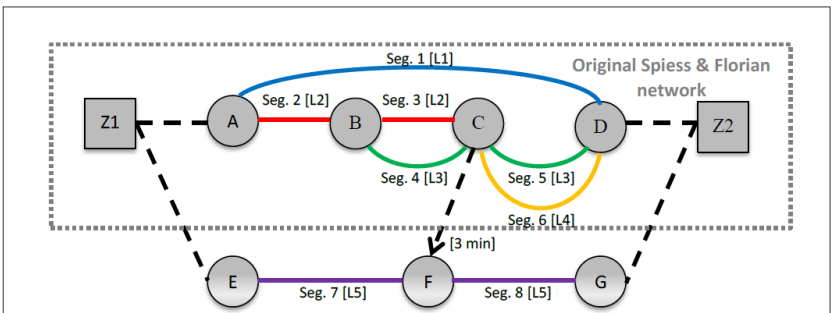
Fig. 10. Explicit capacity constraints' algorithm assumptions in simulations

For both of these, 4 individual O-D demand cases were assigned which should reflect the rising O-D demand conditions in the following stages:

- undersaturated conditions (1600 pax./hour – "LOW" congestion case),
- saturated network state (3200 pax./hour – "MID" congestion case),
- moderately and massively overcrowded conditions (6400 pax./hour – "HIGH" congestion case, and 16000 pax./hour – "V. HIGH" congestion case).

These respective O-D demand values correspond roughly to 50%, 100%, 200% and 500% of generated line capacity (per hour) combined for initial 3 line segments (L1, L2 and L5) departing from the origin. Total simulation run time is 120 minutes: service supply is generated during the whole 120 minutes, whereas passenger demand is assigned after initial 30 minutes and is generated within the next 60 minutes.

Simulations performed on a sample network reveal that both modelling approaches produce different assignment output as a consequence of rising O-D passenger volumes, with respect to each individual (described above) category of passenger overcrowding effects. Starting with the inclusion of physical capacity limits, the implicit approach has relatively more limited impact on output network performance: in aggregate terms, average journey times increase from 21.7 mins ("LOW" congestion case) to just 24.3 mins ("V. HIGH" congestion case) (Fig. 12). These changes in journey times are pretty much minimal and can be merely attributed to the relative shifts in O-D path choices (i.e. paths with longer in-vehicle travel times become somewhat more attractive), but they do not reflect any changes in waiting times - which remain virtually constant (or even decrease slightly) in the event of massive passenger congestion. This stands in stark contrast to the explicit constraints' algorithm which reveals much more significant changes in travel times: as a consequence of rising passenger congestion, average journey times increase from 31.4 mins ("LOW" congestion case) to 63.7 mins ("V. HIGH" congestion case). Here, the average in-vehicle travel times remain constant, but a significant surge in waiting times takes place now due to congestion-induced queuing phenomena at stops, which are evidently captured by the explicit algorithm: as O-D demand volume exceeds the system capacity, a rising share of passengers is denied the boarding and becomes increasingly delayed as they try to reach the destination.



Segment	Line	Out node – In node	Average headways (min)	Travel time (min)	Passenger load in S&F assignment
1	L1 (blue)	A-D	6	25	0.50
2	L2 (red)	A-B	6	7	0.50
3		B-C		6	0.50
4	L3 (green)	B-C	15	4	0
5		C-D		4	0.08
6	L4 (yellow)	C-D	3	10	0.42
7	L5 (purple)	E-F	5	20	-
8		F-G		7	-

Fig. 11. Spiess-Florian extended network (SF ENet) -topology of sample bus transport network used in simulation works (source: Fonzone and Schmoecker, 2015).

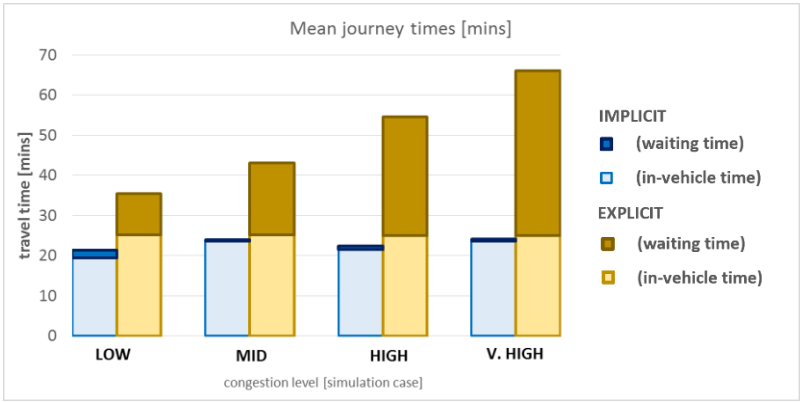


Fig. 12. Simulation results – substantial differences in mean journey times between the explicit and implicit algorithms

In the implicit approach, the effects of arising congestion are described by the travel cost penalty imposed by the SBB function: it reflects the travellers' willingness to shift towards less-crowded connections, but does not account for strict denial-of-boarding: in the end, 100% of travellers will reach the destination successfully and the whole O-D demand volume would be redistributed within the whole 2-hour simulation period to earlier or later departures, even if it implies volume-to-capacity ratio values reaching up to 500% on individual line segments (Fig. 13).

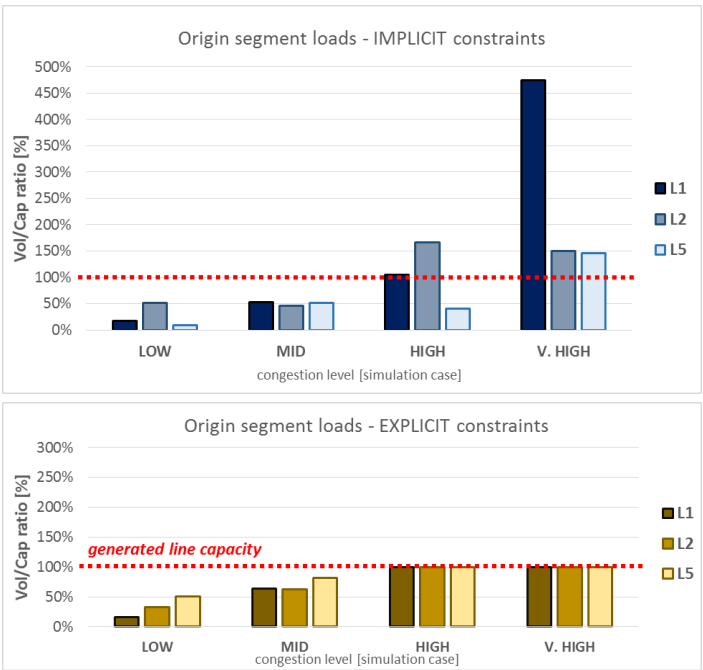


Fig. 13. Results – differences in origin segments' (L1, L2 and L5) Vol/Cap ratios - plotted against the generated line capacity threshold

In contrast, in the explicit approach a strict denial-of-boarding principle is observed for every additional passenger beyond the capacity limit: volume-to-capacity ratio will never exceed 100%, and travellers would have to wait for the ensuing service runs which will have spare on-board capacity. Consequently, the probability-of-arrival at the destination decreases sharply as overcrowding develops in the SF ENet: for "HIGH" and "V. HIGH" congestion cases, 46% and 71% of travellers respectively will not make it to the destination after 120 minutes of simulation run time, and will still remain stranded somewhere in the network (Fig. 14).

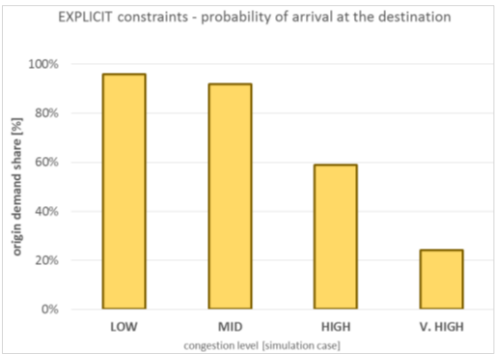


Fig. 14. Results - rising failure-to-arrive probability at the destination in the explicit approach, as a consequence of increasing network congestion

An important remark regarding the mesoscopic-based (explicit) algorithm performance should be made here, which is related to distinct assumptions utilised in the probabilistic discrete choice algorithm. Each time (i.e. at each instance) the traveller makes a travel decision, each alternative he (she) considers in the O-D choice set is described with a non-zero probability – thus, he (she) will *most likely – but not necessarily* – choose the O-D alternative with the highest utility value. Simulation works assumed a default MNL *theta* parameter value of 0.50 – which should be in practice properly calibrated (i.e. most likely, increased) to match the expected probability rate of rational choice behaviour. This comprises a significantly distinct feature of mesoscopic-based algorithm assumptions – and therefore, the exact travel time values should not be interpreted in absolute terms (e.g. in comparison to macroscopic-based algorithm) but rather used to observe relative changes as a consequence of system overcrowding. This is also the reason behind a non-zero failure-to-arrive probability rate (at the destination) even in low congestion scenarios; in higher congestion levels, the additional rises in this probability rate can be directly attributed to the implications of overcrowding-induced phenomena.

A major difference in the assignment output concerns the replication of demand-supply interactions, i.e. the feedback effect between passenger congestion and service performance. This cannot be captured within the implicit approach, where both service run times and dwell times remain unaffected despite the passenger overcrowding – regardless of all the simulation cases. However, it is of utmost importance in case of explicit approach, where mutual dynamic developments on-going in the congested network have profound implications both on the demand (passengers') and the supply (services') side. A significant growth in dwell times can be evidently observed for individual vehicle runs as passenger boarding and alighting flows increase at the stops, which result in up to 50% longer total run times of bus trips in the SF ENet. The demand-supply feedback loop is perhaps best demonstrated when plotting dwelling flows against service headways for consecutive vehicle runs (Fig. 15): it shows that service headways are likely to deviate from their nominal values when fluctuations in dwelling flows grow higher. Importantly, the biggest headway deviation values are correlated not with the extreme demand magnitude - but principally with the *extreme demand variance*: the biggest "bumps" in line headways tend to overlap with the highest "bumps" in dwelling flows. This is a characteristic feature of the on-going bus bunching effect (described above), which in highly overcrowded simulation cases ("HIGH" and "V. HIGH" cases) becomes a self-sustaining phenomenon, reflecting that the network performance falls out of stability state - and will only diminish in the final 30 minutes of simulation period once O-D demand generation ceases and the SF ENet finally "recovers" from massive congestion.

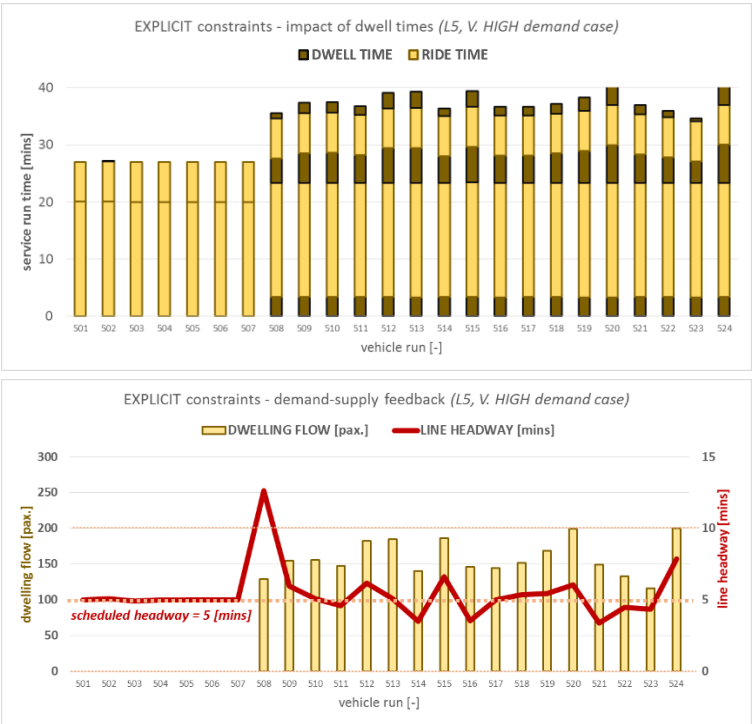


Fig. 15. Results - mutual demand-supply interactions captured in the explicit approach: sample effects on service run times (top) – up to 50% longer service times, and headway deviations (bottom), induced by passenger flows

The differences in the observable assignment output can be attributed to the assumed aggregation level within the simulation algorithm. The implicit approach operates on a macroscopic level, where transport demand and transport supply systems can only be traced in terms of aggregate flows and link segments for the whole assignment period: a more exact examination of service run times' or journey times' distribution is not possible within the scope of this algorithm, and output network performance is principally measurable with average (aggregate) indicator rates. The explicit approach assumes a more disaggregate representation both on the demand as well as the supply side, and thus enables to observe much more detailed output for each individual component of the transport system – i.e. journey times of individual travellers, and service run times of individual vehicle runs. This allows us to reproduce an interesting passenger arrival pattern at the destination, which also mimics the demand-supply feedback interaction: for higher congestion cases, the rising bus bunching effect eventually induces a **"passenger bunching" pattern**, with O-D demand arrivals becoming more concentrated ("bunched") due to system capacity bottlenecks (Fig. 16).

Finally, distribution of path choice patterns also exposes substantial differences between the two assignment algorithms, as seen on the example of path choice shares between 3 line segments at the origin (L1, L2 and L5) (Fig. 17). In the implicit approach, the path choice formula reflects the discomfort cost penalty already for low and moderate crowding conditions. Thus, a substantial shift can be observed when congestion rises from the "LOW" to the "MID" case: the O-D demand becomes pretty much equally distributed between the 3 segments, and for each of them the volume-to-capacity ratio stabilises between 46% to 52%.

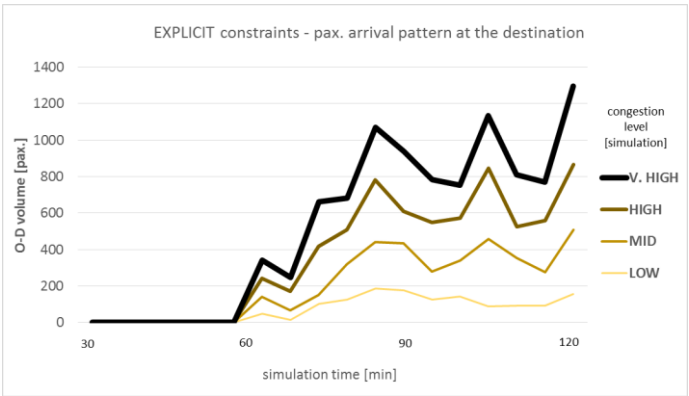


Fig. 16. Results - implications of demand-supply feedback in the explicit approach: fluctuations in service performance (bus bunching) eventually influence the overall pass. arrival ("pass. bunching") pattern at the destination.

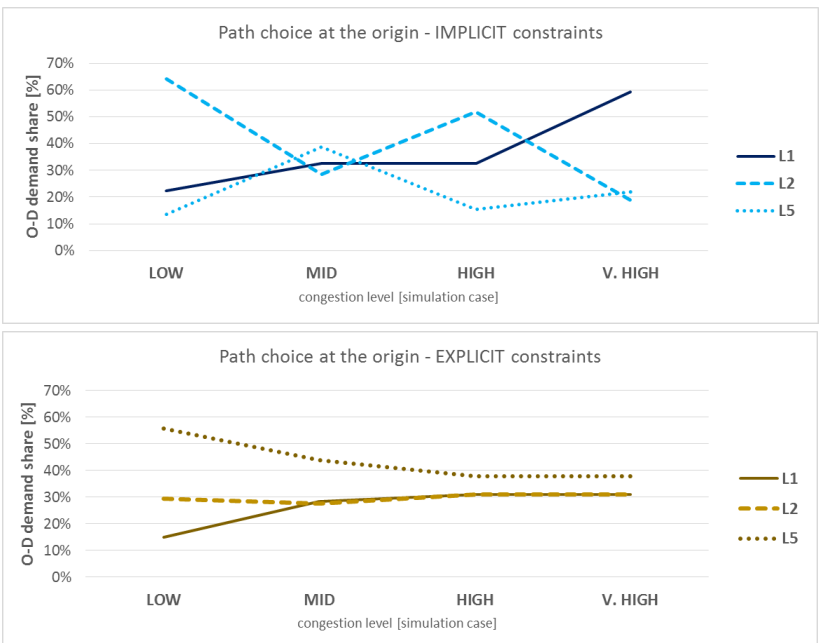


Fig. 17 Results - arising differences in path choice at the origin between the implicit (top) and explicit (bottom) approaches

However, for further ("HIGH" and "V. HIGH") congestion cases no consistent path choice pattern can be derived or explained: the O-D demand shares alternately jump up or drop down, suggesting that the network output could not reach a stable (equilibrium) solution - the small-scale SF ENet with its simple topology becomes simply a few times more overloaded than its generated capacity rate. In the explicit approach, no discomfort cost penalty was included in the path cost formula yet, and the

output path shares reflect merely the *explicit* impact of line segments' capacity limits. This actually produces a pretty much consistent picture of evolving demand distribution pattern: in the relatively less crowded simulation cases, a bulk of O-D demand share is concentrated along the most attractive L5 line segment (56% in the "LOW" congestion case), whereas increasingly congested conditions in the SF ENet result in a more even utilisation (distribution) of the available system capacity and hitherto less attractive O-D routes: in the "HIGH" congestion case, the L5 patronage rate drops to 38% whereas both the L1 and L2 patronage rates reach 31%.

6. Evaluation and conclusions

The objective of this paper was to discuss the incorporation of passenger congestion (overcrowding) effects in public transport assignment models. The first part of this work, being based on a comprehensive literature review, aimed to outline main categories of passenger overcrowding effects, which should be accounted for in the assignment output – i.e. the inclusion of physical capacity constraints (limits); the feedback effect between passenger demand and service supply performance (e.g. the well-known bus bunching effect); and the feedback effect on passenger discomfort (travel cost). Then, a brief presentation of sample case studies followed up, which demonstrated that the inclusion of public transport capacity constraints might significantly affect the actual assignment output and final analysis results – with implications not only for the path (route) choice stage, but also going back to the mode choice and long-term demand adaptation process (e.g. departure time choice). Further on, the simulation part of this paper aimed to examine the replication of passenger overcrowding effects on a small-scale, sample transport network for 2 distinct modelling approaches to public transport capacity constraints – i.e. macroscopic and mesoscopic assignment models, implemented respectively in the PTV VISUM and BusMezzo algorithms.

Both assignment algorithms can reproduce passenger overcrowding effects in a substantially different manner, and will consequently yield quite distinct assignment output (Table 1). The macroscopic algorithm assumes an implicit, simplified approach to modelling capacity constraints of public transport system – i.e. in the form of VDF-based procedure. The increasing travel cost penalty aims to reflect the two-tier effect of arising congestion on passenger path choices – i.e. shifts due to travel discomfort (in low-congested conditions) and demand outflow towards alternative connections (once volume exceeds capacity). As congestion arises in the network, the implicit approach tends to redistribute the O-D demand towards available system capacity and hitherto less attractive O-D paths (routes) – though a certain drawback of this algorithm is that for massive congestion levels, the assignment procedure might not actually converge to a stable (equilibrium) solution and would produce erroneous path choices. However, the implicit-based algorithm does not observe the exact capacity limits of travel supply – eventually 100% of O-D demand will be assigned to the network - nor does it capture the resultant queuing phenomena. Perhaps more importantly, the mutual interaction between transport demand vs. transport supply performance, induced by passenger congestion phenomena, is also missing.

Table 1. Summary - inclusion of 3 distinguished categories of passenger overcrowding effects in the 2 analysed modelling algorithms

Modelling approach:	**IMPACT OF PUBLIC TRANSPORT CONGESTION (PASSENGER OVERCROWDING) EFFECTS – SUMMARY:**		
	Physical capacity constraints (limits)	**Feedback on demand-supply performance**	**Feedback on passenger discomfort**
MACROSCOPIC, timetable-based (PTV VISUM)	IMPLICIT	no*	yes
MESOSCOPIC, simulation-based (BusMezzo)	EXPLICIT	yes	yes**

* limited functionality (i.e. simplified elongation of service run times) only
** not available yet at the time of this research work

The mesoscopic algorithm utilises a more detailed representation of transport system both on demand and supply sides, with a more explicit, detailed approach to modelling capacity constraints – i.e. by observing exact capacity limits of public transport vehicle runs and the ensuing passengers' sequential travel choices. Passengers experience strict denial-of-boarding once a vehicle becomes overcrowded, and as a result the queuing phenomena occur at stops. The passengers' arrival probability is thus heavily influenced by the capacity restraint regime of transport system: the excessive O-D demand share will not arrive yet at the destination if volume exceeds (generated) system capacity. Importantly, the explicit approach enables to reproduce much more dynamic demand-supply interactions, in the form of the on-going feedback between passenger flows and service performance. Since dwell times depend mutually on dwelling flows, a clear-cut, developing bus bunching effect can be traced for individual (consecutive) vehicle runs which is characteristic for congested public transport networks. Resultant path choices reflect shifts both due to available network capacity and current service performance: no feedback effect on path cost (discomfort) was tested yet at the time of this research work, but this can now also be incorporated in the mesoscopic-based algorithm.

Based on the summarised state-of-the-art research works, as well as own simulation works, it seems conceivable that the implicit (macroscopic) modelling approach can be used to model the impact of passenger congestion on path choices in a simplified manner - though a more accurate and reliable representation of the congestion-induced effects is feasible only with an explicit (mesoscopic) modelling approach. The implicit approach is more commonly implemented in the state-of-the-practice assignment models and can be already utilised to replicate a certain (limited) overcrowding impact on route choices and modal shifts in city-scale transport models; the explicit approach remains less developed and often constrained to individual application studies, but comprises a more promising and comprehensive method of representing the whole complexity of public transport congestion effects. Future research works should involve a more specific investigation of model calibration and validation, as well as comparison of assignment output on a city-scale, multimodal transport model – not only just in terms of induced shifts in path choices, but also the far-reaching influence on modal choices, temporal choices and long-term demand adaptation processes (Drabicki et al., 2016) - so as to assess the overall effectiveness of both implicit and explicit algorithms in modelling the public transport network capacity constraints.

References

[1] ATTANUCCI, J. , 2010. Public Transportation Systems, Lecture 14. Lecture presented at the Massachusetts Institute of Technology, Cambridge, MA. Available from: https://ocw.mit.edu/courses/civil-and-environmental-engineering/1-258j-public-transportation-systems-spring-2010/lecture-notes/MIT1_258JS10_lec14.pdf [Accessed 21st May 2017]

[2] BARRY, J., 2015. London's Bus Service – monitoring satisfaction. In *Smart Public Transport Conference*, Warsaw, Poland.

[3] BATARCE, M., MUÑOZ, J. C., ORTUZAR, J. D., RAVEAU, S., MOJICA, C., & RÍOS, R. A., 2015. Valuing crowding in public transport systems using mixed stated/revealed preferences data: the case of Santiago. In *TRB 94th Annual Meeting Compendium of Papers*, Washington DC.

[4] BĄK, R., 2010. Simulation model of the bus stop. *Archives of Transport*, 22(1), 5-25.

[5] BRANSTON, D., 1976. Link capacity functions: A review. *Transportation Research*, 10(4), 223-236.

[6] CASCETTA, E., 2013. Transportation systems engineering: theory and methods (Vol. 49). *Springer Science & Business Media*

[7] CATS, O., 2011. Dynamic Modelling of Transit Operations and Passenger Decisions. Doctoral thesis in Transport Science with specialisation in Transport Systems. KTH – Royal Institute of Technology, Stockholm, Sweden.

[8] CATS, O., WEST, J., & ELIASSON, J., 2014. Appraisal of increased public transport capacity. In *hEART Conference 2014* in Leeds, UK.

[9] DRABICKI, A., 2015. Incorporation of public transport overcrowding effects on passenger path choices in transit assignment. Master

Thesis. Cracow University of Technology, Krakow, Poland.

[10] DRABICKI, A., KUCHARSKI, R., & SZARATA, A., 2016. Zastosowanie ograniczeń przepustowości sieci transportu publicznego w makroskopowym rozkładzie ruchu. (ENG.: Incorporation of public transport network capacity constraints in macroscopic trip assignment model) *Transport Miejski i Regionalny*, 08/2016., Krakow, Poland.

[11] FONZONE, A., & SCHMÖCKER, J. D., 2014. Effects of transit real-time information usage strategies. *Transportation Research Record: Journal of the Transportation Research Board*, (2417), 121-129

[12] FONZONE, A., SCHMÖCKER, J. D., & LIU, R., 2015. A model of bus bunching under reliability-based passenger arrival patterns. *Transportation Research Part C: Emerging Technologies*, 59, 164-182.

[13] GENTILE, G., & NOEKEL, K. (eds.), 2016. Modelling public transport passenger flows in the era of intelligent transport systems. *Gewerbestrasse: Springer International Publishing*.

[14] HARTL, M., 2013. Route choice in macroscopic and microscopic assignment models for public transport. Master thesis. Universitaet Stuttgart, Germany.

[15] HORBACHOV, P., NAUMOV, V., KOLII, O., 2015. Estimation of the bus delay at the stopping point on the base of traffic parameters. *Archives of Transport*, 35(3), 15-25.

[16] KIM, J. K., LEE, B., & OH, S., 2009. Passenger choice models for analysis of impacts of real-time bus information on crowdedness. *Transportation Research Record: Journal of the Transportation Research Board*, (2112), 119-126.

[17] LEURENT, F., 2009. On seat congestion, passenger comfort and route choice in urban transit: a network equilibrium assignment model with application to Paris. In *Annual Meeting of the Transportation Research Board Session Transit Capacity and Quality of Service* (pp. TRB-09). TRB.

[18] LIEBERHERR, J., & PRITSCHER, E., 2012. Capacity-restraint railway transport assignment at SBB-Passenger. In *Proceedings of the 12th Swiss Transport Research Conference*.

[19] LONDON ASSEMBLY TRANSPORT COMMITTEE REPORT, 2009. The Big Squeeze. Rail overcrowding in London. Report commissioned by the Greater London Authority, UK

[20] LONDON ASSEMBLY TRANSPORT COMMITTEE REPORT, 2009. Too close for comfort. Passengers' experiences of the London Underground. Report commissioned by the Greater London Authority, UK.

[21] KROES, E., KOUWENHOVEN, M., DEBRINCAT, L., & PAUGET, N., 2013. On the Value of Crowding in Public Tansport for Ile-de-France.

[22] MOREIRA-MATIAS, L., FERREIRA, C., GAMA, J., MENDES-MOREIRA, J., & DE SOUSA, J. F., 2012, July. Bus bunching detection by mining sequences of headway deviations. In *Industrial Conference on Data Mining* (pp. 77-91). Springer Berlin Heidelberg.

[23] NEWELL, G. F., & POTTS, R. B., 1964. Maintaining a bus schedule. In *Australian Road Research Board (ARRB) Conference*, 2nd, 1964, Melbourne (Vol. 2, No. 1).

[24] NUZZOLO, A., CRISALLI, U., & ROSATI, L., 2012. A schedule-based assignment model with explicit capacity constraints for congested transit networks. *Transportation Research Part C: Emerging Technologies*, 20(1), 16-33.

[25] PTV AG, 2015. VISUM 15 User Manual. Karlsruhe, Germany.

[26] RUDNICKI, A., 1999. Jakość komunikacji miejskiej. (ENG.: Quality of urban public transport.) In *Zeszyty Naukowo-Techniczne Oddziału Stowarzyszenia Inżynierów i Techników Komunikacji w Krakowie*, (71). Krakow, Poland.

[27] SMALL, K. A., & GOMEZ-IBANEZ, J. A., 1999. Urban transportation. Handbook of regional and urban economics, 3, 1937-1999.

[28] SPIESS, H., & FLORIAN, M., 1989. Optimal strategies: a new assignment model for transit networks. *Transportation Research Part B: Methodological*, 23(2), 83-102.

[29] SZARATA, A., 2013. Modelowanie podróży wzbudzonych oraz tłumionych zmianą stanu infrastruktury transportowej. (ENG.: Modelling of induced and suppressed trips resulting from changes in transport

infrastructure) Cracow University of Technology, Krakow, Poland.
[30] SZARATA, A., 2014. The multimodal approach to the modelling of modal split. *Archives of Transport*, 29(1), 55-63.
[31] TIRACHINI, A., HENSHER, D. A., & ROSE, J. M., 2013. Crowding in public transport systems: effects on users, operation and implications for the estimation of demand. *Transportation research part A: policy and practice*, 53, 36-52.
[32] TRANSPORT FOR LONDON, 2013. Capacity for growth at Camden stations. Presented in the London Borough of Camden, 18th September 2013, Greater London, UK.
[33] VAN OORT, N., DROST, M., BRANDS, T., & YAP, M., 2015. Data-driven public transport ridership prediction approach including comfort aspects. *In 13th CASPT Conference*, Rotterdam, The Netherlands.
[34] WHELAN, G. A., & CROCKETT, J., 2009, March.. An investigation of the willingness to pay to reduce rail overcrowding. In *International Choice Modelling Conference 2009*.
[35] ŻOCHOWSKA, R., 2014. Selected issues in modelling of transport flows in congested urban networks. *Archives of Transport*, 29(1), 77-89.

www.czasopisma.pan.pl PAN POLSKA AKADEMIA NAUK www.journals.pan.pl
ARCHIVES OF TRANSPORT *ISSN (print):* *0866-9546*
Volume 43, Issue 3, 2017 *e-ISSN (online):* *2300-8830*
DOI: 10.5604/01.3001.0010.4225

SPEED MANAGEMENT AS A MEASURE TO IMPROVE ROAD SAFETY ON POLISH REGIONAL ROADS

Stanisław Gaca[1], Sylwia Pogodzińska[2]
[1,2] Cracow University of Technology, Department of Civil Engineering, Cracow, Poland
[1]e-mail: sgaca@pk.edu.pl
[2]e-mail: spogodzinska@pk.edu.pl

Abstract: *The article presents the issue of the implementation of speed management measures on regional roads, whose character requires the use of different solutions than those on national roads. The authors briefly described speed management measures, the conditions for their implementation and their effectiveness with reference to environmental conditions and road safety. The further part of the paper presents selected results of the authors' research into the speed on various road segments equipped with different speed management measures. The estimations were made as to the impact of local speed limits and traffic calming measures on drivers' behaviour in free flow conditions. This research found that the introduction of the local speed limits cause reduction in average speed and 85th percentile speed up to 11.9 km/h (14.4%) and 16.3 km/h (16.8%) respectively. These values are averaged in the tested samples. Speed reduction depends strongly on the value of the limit and local circumstances. Despite speed reduction, the share of drivers who do not comply with speed limits was still high and ranged from 43% in the case of a 70 km/h limit, up to 89% for a 40 km/h limit. As far as comprehensive traffic calming measures are concerned, results show decrease in average speed and 85th percentile speed up to 18.1 km/h and 20.8 km/h respectively. For some road segments, however, the values of average speed and 85th percentile speed increased. It confirms that the effectiveness of speed management measures is strongly determined by local circumstances.*

Key words: *speed, speed management, road safety.*

1. Introduction

The previous studies and experience related to the organization of traffic have clearly indicated that it is possible to achieve significant benefits in the area of road safety improvement through the consistent implementation of speed management. However, the effectiveness of speed management measures depends on their proper selection. It is advisable to take into account factors determining the tolerance of the introduced measures on the part of road users. The most important factors from among these include the topography of the road together with the land development and the use of its surroundings. Based on these, a road users subjectively estimates the perceived level of risk. Taking this fact into account, one should pay particular attention to regional roads whose technical standard is more diversified than in the case of the national road network. Regional roads display a diversity of technical classes and function with frequently occurring discrepancies between the road category and its actual function. In addition to this, regional roads are characterized by an uncontrolled accessibility and low geometric parameters of road segments and intersections as compared with drivers' expectations (e.g. narrow lanes, lack of sidewalk). This entails a greater need for speed limits and other speed management measures than in the case of national roads. The way in which these speed limits are introduced ought to be related to the previous assessments of the effectiveness of various speed management measures. With reference to national roads in Poland this types of studies were conducted in 2002-2008 (Gaca, Jamroz, et al., 2003-2008). The issue of local roads has been researched into with this respect on a larger scale since the year 2013. Some part of the research is carried out as a joint project of the Cracow University of Technology and the Gdańsk University of Technology (Jamroz, Gaca, et al., 2013).

One of the reasons behind the research into speed limits on regional and local roads was to find out to what extent the reduced technical parameters of roads and their diversified functions affect the level of tolerance for the general and local speed limits. Also, an important aim of the study was to fill in the

gap in the knowledge on the effectiveness of selected speed reduction measures on local roads presenting lower technical standards. With a wider knowledge on this subject, it will be possible to prepare a more rational implementation of these measures in Poland.

The paper describes selected results of the research into speed limits. The results are important in the formulation of rules governing the implementation of speed management on local roads. The studies were carried out with the participation of authors as a part of national guidelines for speed management (Gaca, Kieć, et al., 2016). The description of the research results was preceded by a review of the most important experiences relative to speed management measures with a particular focus on the way in which they are selected.

2. Speed management measures and their implementation

In the Polish practice of engineering, the speed management is most often understood as introducing speed limits. Therefore, the authors see it purposeful to introduce a broader notion of speed management understood as a set of activities aimed at establishing reasonable speed limits and influencing the actual speed of vehicles through planning, infrastructure and traffic organisation solutions, as well as supervision, education and advanced technologies. Its primary purpose is to achieve a status quo in which the speed of vehicles shall be adapted to the conditions of traffic and road as well as could be considered potentially safe. In addition to this, correct speed management leads to reducing road noise level road and air pollutant emissions.

Speed management consists of activities from the following areas:

- engineering – designing a road infrastructure with parameters facilitating the selection of appropriate speed, the use of physical measures regulating the speed of vehicles, introducing reasonable speed limits,
- supervision – monitoring of drivers' compliance with regulations and speed limits in force.
- education – informing drivers of the impact of speed on road safety, increasing drivers' awareness in terms of the introduced speed limit reduction measures.
- emergency services – enabling emergency services to reach the accident site as fast as possible.
- Speed management should be taken into account not only at the stage of the planning and designing of road infrastructure, but also when it's already in use.
- Among the many engineering measures affecting the vehicles' speed reduction, the following can be listed:
- zone and local speed limits using road signs,
- optical reduction of the width of the lane with the use of a horizontal marking,
- physical narrowing of the road cross-section (one or two-sided),
- traffic islands and a pedestrian refuge,
- raised intersections and pedestrian crossings,
- speed bumps and speed humps,
- horizontal deflection, chicanes,
- converting junctions into roundabouts and mini roundabouts,
- deflecting the vehicle trajectory by the approach to the intersection,
- vibro-acoustic marking.

Only the most beneficial measures, in terms of effectiveness and application costs, tend to be chosen from the group of potential measures suitable for practical application. However, it should be noted that the effectiveness of particular solutions can be strongly determined by local circumstances.

An important tool used in speed management are the rules for determining speed limits, which may be:

- general limits (based on the assumption that roads with similar characteristics can be safely used under ideal conditions with certain maximum speeds),
- zone or local speed limits, most often resulting from the increased risk of accidents in place where the speed limit is introduced.
- The key factors to be taken into account when deciding on zone or local speed limits are (Austroads, 2014):
- accident data (causes and types of accidents, frequency of occurrence, severity of accidents in conjunction with the vehicle speed),
- the geometry of the road and its equipment (width, the field of visibility, curves, intersections, accessibility, barriers etc.),

- road function (arterial roads, collector roads, local roads),
- road users (including the occurrence and size of the pedestrian and cyclist flow),
- the current speed limit,
- actual vehicles speed,
- road environment (including the intensity of the development of the road environment, the potential impact of traffic on inhabitants, including noise, air pollution, separation of the communities, the density of access points to the road),
- the opinion of the local community (inhabitants should have the opportunity to express their concerns and preferences for lower speed limits and their position should be considered).

As an example of the procedure for setting the speed limits, let us present the Australian method, in which the determination of the relevant speed limit is performed in two steps (Austroads, 2010). Step 1 consists in determining the speed limit on the basis of the characteristics of the road (municipal/suburban, one-lane/two-lane, function, flat/hilly/mountainous area, etc.). Step 1 results in the so-called initial speed limit. In step 2, on the basis of the initial speed limit and the appropriate profile of the safe speed limit, the appropriate speed limit is established. Safe speed profiles allow for taking into account road features influencing the likelihood of the accident occurrence and its severity (e.g. the presence of separators dividing lanes going in opposite directions, side obstacles in the road environment). In the case where the restriction resulting from the characteristics of the road is higher than the limit determined by the characteristics of the road, the speed limit at a given road can be increased.

Another method of determining the appropriate speed limit is the approach aiming at the minimisation of the accident impact. Widely described in the foreign literature, this method consists in establishing speed limits based on the biomechanical tolerance i.e. tolerance of the human body to automobile collision impact. The main challenge in this case is the management of the collision force so that no user is exposed to forces that can cause death or serious injury. Table 1 presents a list of allowed speed values in the approach favouring the minimisation of crash effects for various crash types.

Table 1. Biomechanical tolerance of the human body in the event of crash (Austroads, 2005)

Type of the crash	Biomechanical tolerance [km/h]
car/pedestrian	20-30
car/motorcycle	20-30
car/tree or pole (side impact)	30-40
car/car (side impact)	50
car/car (head-on)	70

Apart from establishing the relevant speed limit in a given area, it is equally important to specify the length of the road segment with speed limit. Below there are a few examples of specifying a particular length taken from abroad practice:

- In the USA, in Massachusetts and Ohio it is recommended that the length of a segment with a particular speed limit be equal to at least 0.8 km. In Texas, in turn, the length of the transition zone (buffer zone for the progressive reduction of speed on a suburban road to the speed limit effective in a given locality) must be no less than 0.3 km. Near a school, such zone can have the length of 60-90 m. In Alaska, the length of the speed limit zone shall be determined on the basis of the distance which a vehicle will travel within 25 seconds with the maximum speed determined for this zone (FHWA, 2012a);
- In Canada, the minimum length of the speed limit zone cannot be shorter than 0.5 km (FHWA, 2012a);
- In the UK, in order to avoid too many changes in speed limits on roads, it is assumed that the minimum length of the speed limit segment should be no less than 600 m. With lower speed limits, 400 m is admissible, whereas in the case of access roads or roads with the speed restriction to 30km/h – 300 m (DOT, 2013);
- On two-lane rural roads in Ireland, the speed limit should cover the minimum distance of 3 km. Also, no more than two changes in limits should be introduced on a road segment of 10 km. If the distance between the built-up areas is small (5 km or less), it is appropriate to apply a single speed limit on the road connecting them (DTTAS, 2015).

The change in the value of the speed limit should occur near the place where there is a significant

change in the land development or where the road parameters significantly change.

Irrespective of the method of determining the speed limits and the length of the segment on which they will be effective, particular attention should be paid to the perception of the road by drivers and their expectations as to the speed possible to develop at a chosen road segment. If the driver does not acknowledge the need for a speed restriction, they will not obey it. The conclusion to be drawn is that, e.g. speed restrictions aiming at the reduction of air pollution, in the absence of other "visible" reasons for this restriction, do not always produce the intended effect.

Therefore, before introducing local speed limits, it is widely recommended to introduce other solutions affecting the improvement of road safety. In some cases, the construction of e.g. a cycling path or a sidewalk can more effectively improve the safety of vulnerable road users than a speed restriction on a short segment.

The problem of exceeding the permissible speed limit must be analysed in conjunction with the knowledge of a number of factors determining the drivers' speed. There are numerous works devoted to these issues and their synthesis can be found in (Gaca, 2002; Gaca, Kieć, 2005, Martens, Comte & Kaptein, 1997; Szczuraszek, 2008, Ahie, Charlton & Starkey, 2015, Gaca, Kieć, 2015). In order to quantify the impact of the characteristics of the roads and their surroundings on the speed parameters, regression models are built, the examples of which are given below:

- The estimation of the average speed of vehicles on the suburban road (speed limit of 50 km/h, speed measurements made in the daytime) (Gaca, Kieć, 2015):

$$V_{av} = 70.07 - 1.96 \cdot L - 1.83 \cdot GS + 0.319 \cdot GZ_{50} + 0.169 \cdot T + 0.140 \cdot LZ + 3.55 \cdot C1 + 0.81 \cdot C2 - 4.36 \cdot C3 \quad (1)$$

- The estimation of the average speed of vehicles on the approach to the horizontal curve on two-lane rural road (Jessen, et al., 2001):

$$V_{av} = 55.0 + 0.5 \cdot VL - 0.00148 \cdot AADT \quad (2)$$

- The model for the estimation of the average speed on urban road with a limit of 50 km/h (Schüller, 2010):

$$V_{av} = 48.75 + 1.31 \cdot Ps + 0.88 \cdot \ln(Lskp) + 3.0 \cdot \ln(Bp) + 6.86 \cdot P2x2 - 4.99 \cdot Fh - 2.39 \cdot Fm + 2.74 \cdot Fa + 6.39 \cdot Fb - 1.62 \cdot Stj \quad (3)$$

where:

- V_{av} – average speed of vehicles in free flow in the daytime [km/h],
- L – length of road segment [km],
- GS – density of intersections [number/1 km],
- GZ_{50} – density of development at a distance of 50 m away from the road [%]
- T – share of through traffic (%),
- LZ – average distance between development and the edge of the road (m),
- C1, C2, C3 – the symbol of the road cross-section type (respectively: with bitumic shoulders, ground shoulders and sidewalks). Variable assuming value 1 in the formula (1) if a given cross-section occurs or 0 if it does not,
- V_{85} – 85th percentile speed [km/h],
- VL – speed limit on road [km/h],
- AADT – average annual dly traffic,
- Ps – type of street in network structure (ring road – assumes the value of 0, radial road – assumes the value of 1.0),
- Lskp – length of the segment between the give-way intersections [km],
- Bp – width of the lane taking into account an adjacent bike lane, if it occurs.
- P2x2 – multilane road cross-section (assumes the value of 1.0 if it occurs and 0 if it does not),
- Fh, Fm, Fa, Fb – symbols used to describe the functions of the street and its surroundings, Fh – dominant commercial function, a city centre; Fm – mixed functions; Fa – no dominant functions, unilaterally development; Fb – no dominant functions, no development (variables assuming the value of 1,0 if a given case occurs or 0 if it does not).
- Stj – condition of the road – a variable equals 1.0 in the case of a poor pavement condition and a cobbled road. It equals 0 if the road is in good condition.

When taking into account speed models developed in Poland and abroad, the most frequent and statistically significant quantitative and qualitative variables are: road width, intersections density, density of access points, density of pedestrian crossings and bus stops, curvature degree, value of the speed limit, type of the cross-section, type of shoulder, intensity of the development in the road surrounding, road/street function, time of day. What transpires from the cited speed models is the fact that not all of these variables occur in models simultaneously.

The above speed models suggest that with the same speed limit the average speed of vehicles in the free flow differs significantly depending on the characteristics of the road and tend to be higher than the limit prescribed. Thus, various measures are taken in order to enforce on drivers a greater compliance with speed limits and adjusting their speed to the local conditions on the road.

3. The effectiveness of speed management measures

3.1. Speed management and road safety

The effectiveness of speed management measures in the context of improving road safety can be analysed using direct methods (accident data) or indirect methods describing potential threats.

The influence of speed management measure on road safety can be indirectly assessed through estimating the value of the quotient of the assumed average of the expected value of road safety measure (e.g. number of accident) on a road section with the applied traffic calming measure and the average of the expected road safety measure on the control section fitted with no evaluated measure. That factor, named in the Highway Safety Manual as CMF (Crash Modification Factor), is a fundamental indicator of the assessment of the impact of various treatment on road safety in the USA (Crash Modification Factors Clearinghouse, n.d.; HSM, 2010). The selected values of this factor, estimated on the basis of foreign studies, and in relation to the different speed management measures, location (local, regional and national roads) accident types and their severity, are summarised in Table 2.

Estimating the impact of any speed management measure on road safety can also be made using intermediate criteria, for example the change of vehicle speed caused by a particular measure. The legitimacy of adopting speed as an indirect criterion of road safety assessment is confirmed by both foreign and domestic research. A statistical relationship between the speed and road road safety is logical and has been repeatedly proven (Gaca, 2002, Cameron & Elvik, 2010; Elvik, et al., 2009; Gargoum, El-Basyouny, 2016).

In order to estimate the impact of the change in vehicle speeds on the change in road safety, the so-called "power model" can be used (Cameron & Elvik, 2010). It allows to predict the change in the number of accidents and their victims based on the

Table 2. Estimated values of the CMF coefficients with various measures aiming at reducing speed (Crash Modification Factors Clearinghouse, n.d., www.cmfclearinghouse.org)

Measure	Area	*CMF*	Accident type	Accident severity
decreasing the limit by 9 km/h	all	1.17	all	all
decreasing the limit by 16 km/h	all	0.96	all	all
decreasing the limit by 24-32 km/h	all	0.94	all	all
speed bumps	urban, suburban	0.5 - 0.6	all	serious and slightly injured
transverse rumble strips	urban, suburban	0.66	all	all
	urban, suburban	0.64	all	serious and slightly injured
area-wide or corridor-specific traffic calming	urban	0.89 – 0.94	all	serious and slightly injured
raised pedestrian crossings	urban, suburban	0.54 -0.7	all	serious and slightly injured
raised intersections	none	1.05	all	serious and slightly injured
converting intersections into low-speed roundabouts	all	1.099	all	all
	all	0.473	all	crashes
the introduction of edge line lanes on tangent and curve	suburban	0.963	related to the speed	all
raised bike crossing	none	1.09	vehicle - cyclist	serious and slightly injured

knowledge of the difference in average speed "before" and "after" a given measure is applied. To do this, the following formula is used:

$$W_1 = (V_1/V_0)^a \cdot W_0 \quad (4)$$

where:
- W_0 – a selected criterion of road safety in the period before the introduction of the measure,
- W_1 – a selected criterion of road safety in the same period after the introduction of the measure,
- V_0 – average speed before the introduction of the measure [km/h],
- V_1 – average speed after the introduction of the measure [km/h],
- a – parameter of a model whose value may be assumed based on literature or determined individually based on regression analysys.

An important assumption when using relation (4) is that only speed changes between "before" and "after" periods, and other determinants which affect road safety remain the same.

Based on the research described in (Cameron & Elvik, 2010), the following values of parameter *a* in the equation (4) were estimated (Table 3).

By adopting the values of parameter *a* provided above in the case of roads in urban areas, one can calculate, for instance, that lowering the speed limit of 60 km/h to 50 km/h, with 60 km/h as the average speed at the start and the actual reduction in speed by 3 km/h, would decrease the number of accidents with fatalities by 12%, and the overall number of accidents by 6%. By increasing the degree of respect for the introduced limitation and reaching the actual reduction in average speed by 5 km/h, one would achieve a decrease in the number of accidents with fatalities by 21%, and the overall number of accidents by 10%.

Table 4 summarises results of research concerning the impact of the reduction of the existing speed limit on the change in the number of accidents and their victims and on reducing the average speed of vehicles (Austroads, 2010).

The above-cited examples illustrate well the importance of additional measures of enforcing the respect for speed limits by drivers. The lack of such measures means that the actual reduction in average speed after the introduction of the "new" speed limit is usually ca. 1/4 of the difference between the "new" limit and the one previously in force (Elvik, et al., 2009; Gaca & Kieć, 2005; Gaca, Jamroz, et al., 2003-2008).

In accordance with research carried out abroad and described in (FHWA, 2012b), speed bumps are very effective, contributing to the reduction of the average speed of vehicles by ca. 32 km/h, while speed cushion – of ca. 27 km/h. Research conducted in Poland also confirms this effectiveness. Raised pedestrian crossings, which constitute an obstacle similar to linear speed bumps, usually cause a reduction in average speed by 4.0 ÷ 6.5 km/h. It should be noted, however, that the introduction of speed bumps causes the local speed reduction with its increase on the sections between the bumps, which is an undesired phenomenon due to the exhaust emissions (increased number of manoeuvres of acceleration and braking). Therefore, it is appropriate to use complex traffic calming measures causing the effect of an even reduction in speed on a designated segment of a road.

The introduction of the mini roundabouts and small roundabouts instead of regular intersections leads to the average decrease in average speed by 36 km/h and 54 km/h respectively.

In addition to physical traffic calming measures, less restrictive measures can be applied, for example vertical signs with the recommended speed which inform drivers about the speed at which they should be moving in a given area. Using such signs allows to reduce the speed of vehicles by 3.2 ÷ 5 km/h (FHWA, 2012b).

Table 3. The value of parameter a in the "power model" equation (Cameron & Elvik, 2010)

Type of accident	Roads in rural areas		Roads in urban areas	
	Value of factor *a*	Confidence interval 95%	Value of factor *a*	Confidence interval 95%
With fatalities	4.1	2.9 ÷ 5.3	2.6	0.3 ÷ 4.9
With fatalities and serious injured	2.6	-2.7 ÷ 7.9	1.5	0.9 ÷ 2.1
With casualties in total	1.6	0.9 ÷ 2.3	1.2	0.7 ÷ 1.7

Table 4. The impact of the reduction in speed limits on the average speed and on the reduction in the number of accidents (Austroads, 2010)

Country	Speed limit reduction [km/h]	Changes in average speed	Reduction in the number of accidents and its victims
Denmark	from 60 km/h to 50 km/h (local roads)	3-4 km/h	number of fatalities – 24%, serious injured – 7%, slightly injured – 11%
Germany	from 60 km/h to 50 km/h	-	number of accidents – 20%
Australia (New South Wales)	from 60 km/h to 50 km/h (local roads)	0.94km/h	number of accidents with fatalities -45%, with injured– 22%, the total number of accidents – 23%, accidents with pedestrians – 40%
Australia (Victoria)	from 60 km/h to 50 km/h (local roads)	2-3 km/h	number of accidents with fatalities -21%, with serious injured – 3%, with slightly injuries – 16%, with casualties –12%, fatal accidents with pedestrians – 25%, accidents with serious injured pedestrians –40%
Australia (Southern Australia)	from 60 km/h to 50 km/h (local roads)	3.8 km/h	number of accidents with casualties – 20%, number of serious injured –20%, slightly injured from – 23% to – 26%, fatalities – 40%
Australia (Queensland)	from 60 km/h to 50 km/h (local roads)	5 km/h	number of accidents with fatalities –88%, number of accidents with casualties – 23%
Australia (Western Australia)	from 60 km/h to 50 km/h (local roads)	1 km/h	number of accidents with fatalities –21%, number of accidents with pedestrians – 51%
Australia (ACT)	from 60 km/h to 50 km/h (local roads)	–	total number of accidents – 2.1% (statistically insignificant)
Australia (Victoria)	from 60 km/h to 40 km/h (temporarily, on commercial streets)	–	number of accidents with casualties –8%, number of accidents with pedestrians – 17%

The above examples of research results clearly confirm that, due to speed reduction, the most effective speed management measures are the physical traffic calming measures. Despite that, the decision about their introduction should be preceded by an analysis of the effectiveness of the application of other solutions, such as vehicle activated signs, intensive supervision, installing or removing the line separating lanes or allowing parking along the road. It should also be remembered that physical traffic calming measures cannot be used freely, for example on main arteries, or roads often used by emergency services.

3.2. Speed management and environmental aspects

Modern speed management policy aims not only at improving road safety, but also at protecting the environment and health (ETSC, 2008). Speed management can be a very effective method of fighting the problem of excessive CO_2 emissions, because fuel consumption, and thus carbon dioxide emissions, is dependent on the speed of vehicles (Fig. 1).

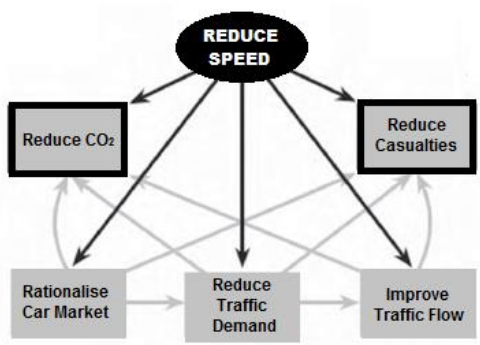


Fig. 1. The impact of the reduction of speed on road safety and reducing CO_2 emissions (ETSC, 2008)

The impact of speed management on environment is widely featured in foreign literature (Austroads, 1996, Bel, Rosell, 2013., COWI & ECN, 2003, ETSC, 2008; Soole, Watson & Fleiter, 2013). The authors of a French government programme showed that total compliance with speed limits would lead to a reduction of CO_2 emissions by 3 million tonnes per

year, which is equivalent to a 2% reduction of emissions of this gas. Even greater benefits were indicated in the research conducted in the Netherlands, in which it was estimated that conducting activities in speed management results in a decrease in carbon dioxide emissions by 4-10%.

The analyses carried out in relation to German motorways suggest that the introduction of 120 km/h and 100 km/h speed limits would reduce carbon emissions by 10% and 20%, respectively. The results of research carried out for the Austrian motorways have shown that the introduction of a 100 km/h speed limit on a 30 km-long section of a motorway leads to a reduction in CO_2 emissions by 11%.

In the United Kingdom, researchers created a model to calculate the emission reduction between 2006 and 2010 for two variants, i.e. for introducing a 110 km/h and 95 km/h speed limits. Studies have shown that for the first scenario, carbon dioxide emissions decreased by ca. 1 million tonnes per year, and for the second one, by 1.88 million tonnes per year.

It is estimated that reducing the speed limit from 100 km/h to 80 km/h on roads in Italy would help to reduce fuel consumption by 387.9 tonnes per year, and to reduce the emissions of CO, PM_{10}, CO_2 and NO_x respectively by 15.3%, 6.4%, 5% and 4.6%. Unlike Italian research, results of similar measurements conducted in Spain showed that reducing the speed limit from 120km/h or 100km/h to 80 km/h on motorways in Barcelona causes increase in NO*x* and PM10 emission by 1.7–3.2% and 5.3–5.9% respectively. This study also suggest that NO*x* and PM10 pollution can be reduced by 7.7–17.1% and 14.5–17.3% respectively after variable speed policy implementation.

The authors of the studies conducted abroad stress that special attention should be paid to the speed of lorries (trucks). The Dutch study shows that the reducing the speed limit for vans and light trucks to 110 km/h on expressways can lead to a reduction of fuel consumption by 5%. In turn, the reduction of vehicle speed by 2 km/h, 5 km/h and 7 km/h results in the reduction of fuel consumption by 0.5% (7 million litres), 2% (27 million litres) and 3.5% (48 million litres) respectively.

Another factor which can have a significant impact on the reduction of CO_2 emissions is the use of modern technologies, such as ISA (Intelligent Speed Assistance). It is estimated that carbon dioxide emissions from vehicles fitted with ISA may drop by 8% in the UK when compared to other vehicles.

Another study carried in the United Kingdom focused on the impact of speed control on fuel consumption and emissions. It has been shown that speed control on roads where speed is reduced to 110 km/h will increase fuel consumption by 70.6 litres/100 km, and on the road segments with the limit of 80 km/h by 18.8 litres/100 km. In addition, the roads segment with speed limits of 110 km/h reported a decrease in CO_2 emissions by 528 kg/month, and on the road segments where the limit is 80 km/h the decrease was by 1376 kg/month.

4. Polish surveys of the effectiveness of speed management measures

Although most of the speed management measures described in point 2 are implemented in Poland, however, the evaluation of their effectiveness has been the subject of limited research only, conducted primarily on national roads and in the cities (Gaca & Kieć, 2005; Gaca, Jamroz, et al., 2003-2008). The general conclusion from these studies is that the drivers of vehicles commonly exceed the permissible speeds determined by the general or local restrictions. The results described in (Gaca & Kieć, 2005; Gaca, Jamroz, et al., 2003-2008) and other test results clearly indicate the need for additional speed management measures, especially on road segments passing through small and medium-sized towns.

In the case of local roads the typical speed management measures include:

- local speed limits – 36.8%,
- locally applied traffic calming measures – 32.3%,
- complex traffic calming measures (median and refuge islands on roads through built-up areas which cause change of trajectory only for heavy vehicles) – 19.4%,
- speed zones – 6.4%,
- designating road segments with intense speed enforcement (along with speed cameras) – 4.0%,
- measures other than those listed above – 1.1%.

These data were obtained as a result of road administration survey research (Gaca, Kieć, et al., 2016).

Taking into account the above-mentioned frequency of applying the speed management measures on local roads, these measures underwent research intended to precede the development of guidelines

for speed management on local roads (Gaca, Kieć, et al., 2016). In subsequent parts of this article, only selected result of speed measurement on 100 road segments and in 24 speed zones were presented.

The basic indicators of the effectiveness of speed management measures are changes in the value of the different criteria of road safety. One can also assess the above-mentioned effectiveness through speed measurements and the evaluation of changes in the driver's behaviour as a response to the applied measure. The expected effect of speed management measures is not only a reduction in the value of average speed, but also a reduction in the share of drivers moving at very high speeds, and the emergence of more uniform behaviour. Therefore, the primary measures of effectiveness are: reduction of average speed, reduction of the value of the 85th percentile speed (V_{85}) and reduction of dispersion in the value of the speed. Their estimation is made by a typical "before and after" or "with and without" studies. The equivalent of "with and without" studies are simultaneous speed measurements conducted on different segments of the road, i.e. in the segment preceding the applied measure, and in the segment where the measure is applied. A group of vehicles in free flow was singled out in the study, which allowed for a better assessment of the driver's reactions to a given measure. A headway greater than 6 sec was assumed as the boundary value for free flow speed. In this case driver has the freedom to make decisions on the selection of speed. As measurement techniques, the following were used: measurements using pneumatic tubes, video technique and measurements using manual devices. The research was carried out in similar weather conditions and during a day. The duration of measurements and the obtained sample sizes met the requirements of mathematical statistics.

In Poland, the effectiveness of local speed limits on lower classes roads is relatively not very well identified. The following describes the selected results of the speed measurements on the segments of roads with local speed limits of 40, 50, 60 and 70 km/h. Pooled results of speed measurement are provided in Table 5.

By analysing a group of vehicles in free flow in road segments with local speed limits, it was found that:

- a very large group of vehicles were moving at a speed greater than the limit. The share of drivers who do not comply with applicable restrictions ranged from 43% in the case of a 70 km/h limit, up to 89% for a 40 km/h limit;
- the introduction of the local speed limit changes the value of the average speed and V85 on average by 4.4 ÷ 11.9 km/h and 5.8 ÷ 16.3 km/h respectively, depending on the value of the limit. This means a change in the average speed and V85 respectively by 6.8 ÷ 14.4% and 7.7 ÷ 16.8%;
- having regard to the results of the measurements in particular road segments, wide variations in reactions of drivers to the introduced restrictions

Table 5. Results of the speed test on road segments with local speed limits (Gaca, S., Kieć, M., Jamroz, K., et al., 2016)

		Section without a speed limit				Section with a speed limit				Difference		
No. of segments	Speed limit [km/h]	Mean speed	85th percentile speed	Coefficient of speed dispersion	Share of drivers exceeding speed limit	Mean speed	85th percentile speed	Coefficient of speed dispersion	Share of drivers exceeding speed limit	Mean speed	85th percentile speed	Coefficient of speed dispersion
		[km/h]	[km/h]	-	[%]	[km/h]	[km/h]	-	[%]	[km/h] [%]	[km/h] [%]	- [%]
8	40	64.4	75.5	0.185	45	60.0	69.7	0.173	89	4.4 6.8%	5.8 7.7%	0.011 6.2%
4	50	69.9	82.1	0.185	47	60.8	70.7	0.179	77	9.1 13.0%	11.4 13.9%	0.005 2.9%
4	60	82.3	97.0	0.174	24	70.4	80.7	0.152	81	11.9 14.4%	16.3 16.8%	0.022 12.5%
9	70	79.9	93.2	0.170	19	70.6	83.3	0.194	43	9.3 11.6%	9.9 10.6%	-0.024 -14%

were found. Depending on the value of speed limits, the average speed on the various segments changed from -5.6 km/h (average speed increase) to 24.7 km/h in relation to the segment without speed restrictions. This means the presence of a strong influence of local factors on the level of tolerance of local speed limits. It is planned to extend the scope of the research to build regression models quantifying the effect of these factors;
- the value of speed V_{85} was higher than the average speed by 11.1 ÷ 14.7 km/h on average in the control section and by 9.7 ÷ 12.7 km/h in the section with speed limits;
- the research recorded a change in the value of the variation coefficient in the road segments with a speed limit in relation to the control sections. The change equalled up to 0.02. In the case of a 70 km/h speed limit, there has been a 0.02 increase in the coefficient of variation in the road segments with this measure, indicating the occurrence of an unexpected effect of the increasing heterogeneity of traffic.

One of the measures of enforcing speed limits are traffic calming measures and the intensive speed enforcement measures. In the case of physical traffic calming measures, among those tested there were complex measures implemented along road segments passing through small localities. Usually, these are "mild" measures designed in a way that allows drivers to pass with the maximum speed of ca. 70 ÷ 80 km/h. Therefore, their effect in urban areas mainly consists of both their psychological impact (drawing attention to the need to reduce speed) and the physical speed reduction for a group of very fast driving vehicles.

Based on tests carried out on 23 road segments passing through localities with complex traffic calming measures, it was found that (Gaca, Kieć, et al., 2016):
- the average speed in free flow in control sections, i.e. before the road segments featuring traffic calming measures, oscillated between 42.8 km/h and 75.4 km/h. In road segments with a 60 km/h speed limit and complex traffic calming measures, this speed ranged from 44.0 km/h to 65.8 km/h;
- the value of speed V_{85} was higher than the average speed by 10.0 km/h on average in the control section and by 8.3 km/h in the section with traffic calming measures;
- the majority of analysed segments with traffic calming measures saw a decrease in average speed and V85 in relation to the control section by 0.1 ÷ 18.1 km/h and 0.6 ÷ 20.8 km/h respectively. For some segments, however, the values of average speed and V85 increased. In total, the tested road segments showed an average decrease in average speed and V85 in free flow by 2.8 km/h and 4.7 km/h respectively;
- complex traffic calming measures on road segments passing through localities have a positive impact on the improvement of traffic homogeneity: the value of the averaged coefficient of speed dispersion in free flow decreased by 0.02 when compared to control segments;
- the share of drivers going over the speed limit in the section with traffic calming measures was 68% and was 11% lower than in the control section.

Preliminary research into the effectiveness of physical traffic calming means in street segments have confirmed the impact of these measures on the reduction of vehicle speeds. Single speed bumps and raised intersections caused a local decrease in speed by 15.7-27.0 km/h. When it came to speed bumps placed consecutively one after another, the object of the study was the average speed in that street segment, which was compared with the speed in the same segment but without speed bumps. The presence of speed bumps resulted in a reduction of the average speed by 20-23 km/h. In the qualitative sense, the quoted examples of random test results fall in line with the results of research by other authors, and they confirm the high efficiency of the physical traffic calming measures.

The above-mentioned examples of studies of the impact of speed management measures on the reduction of the different characteristics of speed indicate a high potential of such management as a means of improving road safety. Although the averaged values of speed reduction of ca. 3 ÷ 15 km/h may seem relatively small, their importance in relation to the improvement of road safety, assessed on the basis of the model described in section 3, is very high. Figure 2 shows the extent to which the change in average speed in a given segment of the road, respectively by 3, 6, 9, 12 or 15 km/h, will affect the reduction (expressed in percent) in the number of accidents with fatalities and serious injured depending on the level of speed before the change.

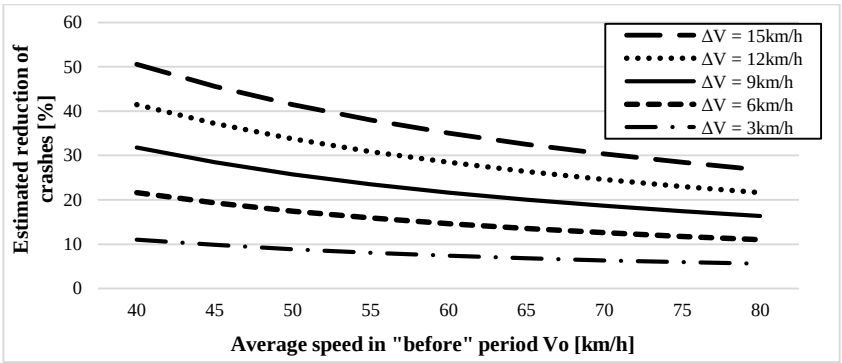
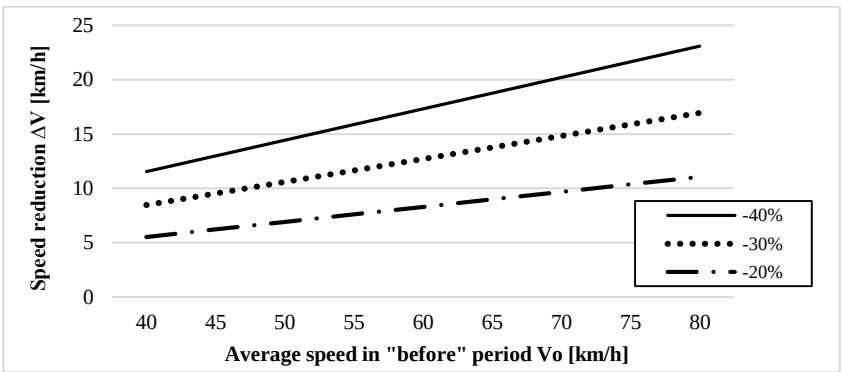
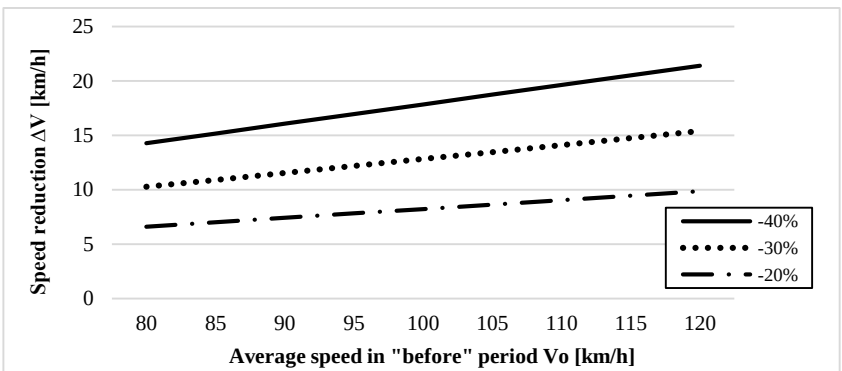


Fig. 2. Estimation of the possible reduction in accidents with fatalities and serious injured depending on the average speed in "before" period and the speed reduction value ΔV, based on the formula (4) (own work).



a)Segments in urban areas



b) Segments in rural areas

Fig. 3. Estimation of the possible reduction in accidents with fatalities and serious injured depending on theaverage speed in "before" period and the speed reduction value ΔV, based on the formula (4) (own work).

The relation shown in Figure 2 indicates a connection between the effect of a potential reduction in the number of accidents and the value of speed "before" the implementation of a specific speed management measure. In addition, this effect is illustrated in Figure 3 which shows the desired speed reduction values in order to obtain the intended value of the reduction in the number of accidents with fatalities and serious injured.
The results of the analyses provided above indicate the need to implement more restrictive speed management measures at side with higher recorded speeds.

5. Discussing the results and the summary

The presented research, in conjunction with other Polish publications, confirm the low level of tolerance of general and local speed limits in Poland. In the case of general speed limits, the share of drivers not adhering to these limitations is significantly larger than in other countries of the European Union (Gaca & Kieć, 2005 Jamroz, Gaca, et al., 2013).
Respecting local speed limits are mainly influenced by the principles of their imposition. If they are not understandable for drivers, then a large role is played by using supervision and additional measures aimed at enforcing the desired behaviour. The general cultural and sociological circumstances, which may vary depending on the region, are also of importance. These circumstances limit the possibility of direct comparison of national research results with the results of research carried out abroad.
The view that the local speed limits are not effective and therefore do not contribute to the improvement of road safety, is quite common. In order to verify this thesis, the research was undertaken. This research found that the reduction in average speed in an area with local speed limits was, on average, 4.4 ÷ 11.9 km/h (6.8% ÷ 13.0%), depending on the value of the limit. These values are lower than expected, but they can potentially cause a reduction in the number of accidents with fatalities and serious injured by ca. 10 ÷ 20%. This is sufficient proof for the legitimacy of the use of local speed limits in areas of increased accident risk. Of course one should aim to improve the degree of compliance with local speed limits, which requires the implementation of more effective speed management measures. Their correct choice requires a better understanding of the determinants of drivers speed choice in conjunction with local circumstances. The results of the presented research clearly indicate the presence of a strong influence of local factors on the level of acceptance of local speed limits.
Research of speed on roads with "mild" traffic calming measures confirmed their low effectiveness, which points to the need for verification of both these solutions and the ways of their implementation. A high effectiveness was, in turn, confirmed, when it comes to physical speed reduction measures in the form of speed bumps and raised intersections. It must be stressed, however, that these measures usually cause only a local decrease in speed.

Acknowledgment
The authors would like to gratefully acknowledge the financial support given by National Road Safety Council in Poland.

References

[1] AUSTROADS LTD., 1996. *Urban speed management in Australia. Sydney*. Report AP-118-96.
[2] AUSTROADS LTD., 2005. *Balance between Harm Reduction and Mobility in Setting Speed Limits: A Feasibility Study.* Report AP-R272-05.
[3] AUSTROADS LTD., 2010. *Infrastructure/Speed Limit Relationship in Relation to Road Safety Outcomes*. Report AP-T141/10.
[4] AUSTROADS LTD., 2014. *Methods for Reducing Speeds on Rural Roads Compendium of Good Practice.* Report AP-R449-14.
[5] BEL, G. & ROSELL, J., 2013. *Effects of the 80 km/h and variable speed limits on air pollution in the metropolitan area of Barcelona*, Transportation Research Part D: Transport and Environment, 23, 90–97.
[6] CAMERON, M.H., & ELVIK, R., 2010. Nilsson's Power Model connecting speed and road trauma: Applicability by road type and alternative models for urban roads. *Accident Analysis and Prevention*, 42, 1908-1915.
[7] COWI & ECN, 2003. International CO2 policy benchmark for the road transport sector.

[8] Crash Modification Factors Clearinghouse (n.d.). Available at: http://www.cmfclearinghouse.org (access date: March 2016)
[9] DOT Department of Transport, 2013. Setting local speed limits. *Department for Transport Circular* ,01.
[10] DTTAS Department of Transport, Tourism and Sport, 2015. *Guidelines for setting and managing speed limits in Ireland*.
[11] ELVIK, R., HØYE, A., VAA, T., & SORENSEN, M., 2009. *The Handbook of Road Safety Measures* (2nd ed.). United Kingdom: Emerald Group Publishing.
[12] ETSC European Transport Safety Council, 2008. *Managing Speed Towards Safe and Sustainable Road Transport.*
[13] FHWA Federal Highway Administration, 2012a. *Methods and Practices for Setting Speed Limits: An Informational.* Report FHWA-SA-12-004.
[14] FHWA Federal Highway Administration, 2012b. *Speed Management: A Manual for Local Rural Roads Owners*. Report FHWA-SA-12-027.
[15] GACA, S., & KIEĆ, M., 2005. Badania reakcji kierujących pojazdami na zmianę ograniczenia prędkości na terenach zabudowy. *Transport Miejski i Regionalny*, 12, 9-14.
[16] GACA, S., KIEĆ, M., 2005. *Models of traffic flows with speed limits.* Archives of Transport, vol. 17 no 2, 15-34
[17] GACA, S., 2002. *Badania prędkości pojazdów i jej wpływu na bezpieczeństwo ruchu drogowego.* Kraków, Zeszyty Naukowe PK - Inżynieria Lądowa, 75.
[18] GACA, S., 2002. *Regression models of accidents and accident rates.* Archives of Transport, vol. 14 no 3, 17-30
[19] GACA, S., JAMROZ, K., et al., 2003-2008. *Analiza wybranych aspektów zachowania użytkowników dróg (Analysis of choosen aspects of road users' behaviour)*. Reports SIGNALCO – FRIL, Krajowa Rada Bezpieczeństwa Ruchu Drogowego.
[20] GACA, S., KIEĆ, M., & ZIELINKIEWICZ, A., 2012. *Identyfikacja determinant bezpieczeństwa ruchu w warunkach nocnych ograniczeń widoczności*. Politechnika Krakowska, Report N509 254437.
[21] GACA, S., KIEĆ, M., 2015. *Research on the impact of road infrastructure on traffic safety*. Chapter in Monograph 483, Recent advances in civil engineering: road and transportation engineering. Politechnika Krakowska, 51-79.
[22] GACA, S., KIEĆ, M., JAMROZ, K., et al., 2016. *Badania skuteczności środków zarządzania prędkością i ich wyniki (Research on effectiveness of speed management measures and its results).* Krajowa Rada Bezpieczeństwa Ruchu Drogowego, Research report. available at: www.krbrd.gov.pl/pl/pozostale.html
[23] GARGOUM, S.A., & EL-BASYOUNY, K., 2016. *Exploring the association between speed and safety: A path analysis approach*. Accident Analysis and Prevention, 93, 32-40.
[24] HSM *Highway Safety Manual*. (1st ed., 2010). Washington DC, AASHTO.
[25] JAMROZ, K., GACA, S., et al., 2013. *Prędkość pojazdów w Polsce w roku 2013 (Vehicles' speed in Poland in 2013)* Krajowa Rada Bezpieczeństwa Ruchu Drogowego, Research report, available at: www.obserwatoriumbrd.pl/resource/38187803-1a05-48c9-ac57-0b9b3e9b6801:JCR
[26] JESSEN, D., SCHURR, K., McCOY, P., & HUFF, R., 2001. *Operating Speed Prediction on Crest Curves of Rural Two-Lane Highways in Nebrask*a. Transportation Research Record: Journal of the Transportation Research Board, 1751, 67–75.
[27] LIV M. AHIE, SAMUEL G. CHARLTON & NICOLA J. STARKEY, 2015. *The role of preference in speed choice*. Transportation Research Part F: Traffic Psychology and Behaviour, 30, 66-73.
[28] MARTENS, M., COMTE, S., & KAPTEIN, N., 1997. *The effects of road design on speed behavior: a literature review.* TNO Human Factors Research Institute, Report TM 97 B021.
[29] SCHÜLLER, H., 2010. *Modelle zur Beschreibung des Geschwindigkeitsverhaltens auf Stadtstraßen und dessen Auswirkungen auf die Verkehrssicherheit auf Grundlage der Straßengestaltung (Speed models for street and influence of speed on road safety with consideration of road parameters)*. Dresden, Schriftenreihe des Instituts für

Verkehrsplanung und Straßenverkehr, ISSN 1432-5500, 12.
[30] SOOLE, D., WATSON, B., & FLEITER, J., 2013. *Effects of average speed enforcement on speed compliance and crashes: a review of the literature*. Accident Analysis and Prevention, 54, 46-56.
[31] SZCZURASZEK, T., 2008. *Prędkość pojazdów w warunkach drogowego ruchu swobodnego (Vehicles' speeds in free flow conditions)* Studia z zakresu inżynierii. Warszawa: PAN KILiW.

ARCHIVES OF TRANSPORT www.czasopisma.pan.pl PAN POLSKA AKADEMIA NAUK www.journals.pan.pl ***ISSN (print):*** *0866-9546*
Volume 43, Issue 3, 2017 ***e-ISSN (online):*** *2300-8830*
DOI: *10.5604/01.3001.0010.4226*

ANALYSIS OF THE INFLUENCE ON EXPRESSWAY SAFETY OF RAMPS

Juan Juan Hu[1,2], Feng Li[2], Bing Han[2], Jinbao Yao[3]

[1] College of Architecture and Civil Engineering, Beijing University of Technology, Beijing, P.R.China

[2]Transport Management Institute, Ministry of Transport of the People's Republic of China, Beijing, P.R.China

[3]School of Civil Engineering, Beijing Jiaotong University, Beijing, China

[3]e-mail: bao_yaojin@163.com

Abstract: *As the very important parts of the expressway system, on and off ramps have a great effect on the operation effect of expressways. Once congestion occurs at on and off ramps, it will directly affect the safety of the expressway and vehicles. So this paper first analyzes the location design of on ramps, the length and the influences on the expressway running state of off ramp downstream traffic state, and establishes the expressway operation models based at on and off ramps. The expressway operation models include off ramp delay model, on ramp delay model, expressway delay model and side road delay model under the influence of the distance between on and off ramp and traffic volume. At the same time, the paper analyzes the connection between the traffic operation state and traffic safety, and establishes the expressway safety model based on congestion degree. Through simulation verification, the impact of on and off ramp on the security of the expressway is analyzed. In the simulation, safety rank division is simulated to verify and safety index of different safety ranks. Finally the paper concludes that the delay caused by traffic of the on and off ramps can decrease the safety of expressway.*

Key words: *expressway safety, on and off ramp, delay, accidents occurrence probability.*

1. Introduction

Due to on and off ramps on expressway, a large number of vehicles access and exit from expressway and the traffic on the outside lane slows to a crawl. Due to the relatively short distance between some ramps, traffic flows between the adjacent ramps will be frequent confluent, diverted and mixed. This is easy to cause conflictions and traffic safety problem (Karoń and Żochowska, 2015). Because there is no conflict point near the ramps, strict management and control measures are not taken generally. Therefore, conflictions of traffic volume near on and off ramps lead to congestions of expressway and side road (Handke, 2010).

Domestic and foreign scholars have done a lot of researches on on and off ramp setting. Leisch (1959) studied interchanges spacing of urban expressway. Al-Kaisy et al. (2002) analyzed the relationship between the ramp traffic flow, traffic capacity, main road traffic flow, length of deceleration lane and lane number in detail. Munoz and Daganzo (2002) described the traffic operation of the bottleneck section of the highway exit upstream in detail. Jayakrishnan et al. (1995) simulated traffic flow of urban road, and proposed that there are many differences between urban road and highway, especially the mutual influences between multiple vehicle types in urban road. Kojima et al. (1995) selected an x-shaped-cross highway as the simulation object to study the mixed behavior of individual vehicle, and analyzed in which part of the weaving section mixed vehicles began to interweave. Research model consisted of three basic equations, which are simple forward movement, car following movement and braking movement.

In addition, Wang et al. (2015) presented Bayesian logistic regression models for single vehicle (SV) and multivehicle (MV) crashes on expressway ramps by using real-time microwave vehicle detection system data, real-time weather data, and ramp geometric information. Qu et al. (2014) aimed to assess the potential crash risks across different traffic lanes (shoulder lane, median lane, and middle lane) near to ramps (before on-ramps, between ramps, and after off-ramps). GAO and GUO (2006) took the common parallel ramp setting as the

research object and put forward five ramp optimization design methods, which are changing location of the ramp, changing lane function of the intersection approach, setting left-turn in long distance, and changing lanes ahead on the side road, etc. Furthermore, Lee and Abdel-Aty (2006) proposed the method of predicting the temporal variation in crash risk on freeway ramps and at the intersections of ramps (the junction of ramps with crossroads). Bared et al. (1999) examined the impact of acceleration and deceleration lane lengths on traffic safety; Chen et al. (2009) evaluated the safety effects of the number and arrangement of lanes on freeway off-ramps; Liu et al. (2009) discussed three types of lane arrangements on freeway sections with closely spaced ramps and their safety effects; Pulugurtha and Bhatt (2010) analyzed the role of geometric characteristics and traffic on crashes in weaving sections.

Researches both at home and abroad showed that the existing researches are mainly about on and off ramps setting, but rarely involves the safety. This paper hopes to propose the relationship between on and off ramp and expressway safety based on the above research results.

2. Expressway running condition analysis on and off ramp

On and off ramp locations on expressway are closely related to urban network structure, and the matching way of off ramp and on ramp is also different. The paper mainly study two typical matching types of on and off ramp on expressway, that are on-off ramp group and off-on ramp group as shown in Figure 1.

The difference between two groups is that for on-off ramp group the distance between off ramp ahead and on ramp behind is fewer than the distance between the off ramp and other ramps, on the contrary, for out-on ramp group the distance between on ramp ahead and off ramp behind is fewer.

The short distance between on and off ramp will lead to congestion of expressway main road and side road, so the paper studies the two matching types. If the distance is long, on and off traffic flow on expressway will not interfere with each other, and the running condition of expressway main road and side road will not be affected. So the long distance condition is not the scope of this study.

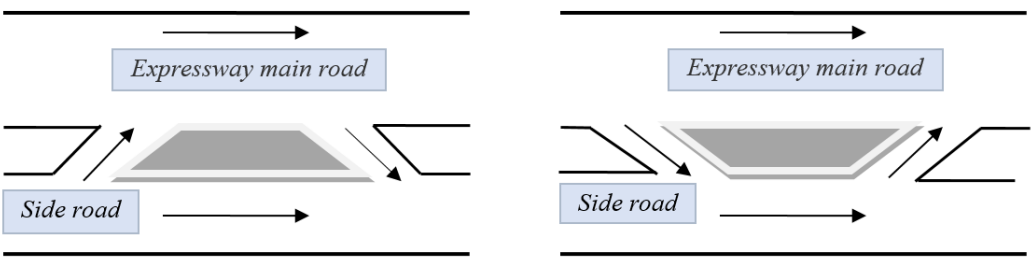
The definition of the distance size between on and off ramp has been reflected in the existing research results, so the paper does not study this problem anymore, but focuses on the influence of on and off ramps on traffic running state of expressway main road and side road.

2.1. On-off ramp group

As shown in Figure 1(a), once off ramp queuing occurs, and ascend to the expressway main road, it not only affects the normal traffic operation of expressway main road and produce corresponding delay, but also affects vehicles of on ramp entering the main road, and causes corresponding delay of side road. Based on the analysis, using critical gap theory and queuing theory, the expressway delay model and side road delay model can be established based on the queue length at on and off ramp.

Model principle is that vehicles enter off ramp based on critical gap theory, this maybe generate certain delay and queue.

According to queuing theory, input process of off ramp traffic is assumed as fixed-length distribution, that is $\lambda = q_c(veh/s)$, because of assumption of large traffic flow.



a) On-off ramp group b) Off-on ramp group

Fig. 1. on and off ramp on expressway

Queuing rule is first come first service, and service station is one lane. Service time distribution is assumed as exponential distribution, that is $\mu = E\left(q_f(t)\right) = \overline{q_f}(veh/s)$, because this paper argues that side road traffic volume is relatively large, critical gap is not uniform. Service intensity is that $\rho = \lambda/\mu$, and queuing model is *D/M/1*.

Then average queue length of off ramp is$L_{s=}\frac{\lambda}{\mu-\lambda}$, average queue vehicles number is $L_q = \rho L_s$

Delay of off ramp is

$$d = W_s + W_q = \frac{L_s}{\lambda} + \frac{L_q}{\lambda} = \frac{1+\rho}{\mu-\lambda} \quad (1)$$

where:

- μ is the average flow rate of the expressway exit ramp, which is the average service rate that side road traffic can serve for the main road traffic. The unit is veh/s.
- $\overline{q_f}$ is the average flow of expressway side road. The unit is veh/s.
- ρ is the service strength for queuing theory, which is the ratio of arrival rate and departure rate of off-ramp traffic flow into side road.
- L_s is the average number of queuing vehicles affected by side road traffic at the off ramp. The unit is *veh*.
- L_q is average number of queuing vehicles waiting for the off ramp. The unit is *veh*. d is the delays of traffic flow at the off ramp. The unit is s.
- W_s is customer stay time in queuing theory, which is the average travel time of all vehicles at the off ramp exiting from the main road. The unit is s.
- W_q is customer waiting time in queuing theory, which is the average waiting time of all vehicles at the off ramp exiting from the main road. The unit is s.

In addition, the maximum queue length of off ramp is also influenced by the signal control on side road. If there is signal control at the downstream intersection of off ramp, and the queue length may ascend to off ramp, forming time of the biggest queue length on off ramp is affected by the signal intersection, If the queue vehicles at downstream signal intersection do not ascend to off ramp, off ramp traffic will flow into side road with critical gap.

Therefore, service time of traffic flow out of off ramp under the influence of the side road signal control change into,

$$\mu = \overline{q_{ff}} = \frac{S\times 3600\times\frac{g}{c}}{3600} = S\times\frac{g}{c}(veh/\mathrm{s}) \quad (2)$$

Therein,

- S is saturation traffic flow rate released in green time at the downstream signal control intersection.
- $\overline{q_{ff}}$ is average traffic of expressway side road affected by traffic signal on side road, the number of vehicles that can be discharged at each cycle at the downstream intersection of the side road is the section traffic volume at the off ramp. The unit is veh/s.
- g is green time at the downstream signal control intersection.
- C is signal cycle at the downstream signal control intersection.

Then average delay of off ramp is,

$$d_c = \frac{1+\rho}{\mu-\lambda} = \frac{\mu+\lambda}{\mu(\mu-\lambda)} = \begin{cases} \frac{q_c+\overline{q_f}}{\overline{q_f}(\overline{q_f}-q_c)}\ q_c(C-g)\times 7+20 \le L_c \\ \frac{q_c+\overline{q_{ff}}}{\overline{q_{ff}}(\overline{q_{ff}}-q_c)}\ q_c((C-g)\times 7+20 > L_c \end{cases} \quad (3)$$

Therein,

- $L_c$is the distance between the downstream signal control intersection and off ramp,
- q_c is traffic volume out of off ramp,

According to average queue length of off ramp, queue dissipation time can be calculated by $t_f = \frac{L_q}{\mu}$. The delay calculation at expressway during the period of time is similar with that of the signal control intersection in one cycle, and the red time is d_c, that is the forming maximum queue length time of traffic flow out of off ramp, and green time is $\frac{q_k d_c}{(S-q_k)}$, that is the dissipation time of traffic flow out of off ramp.

According to steady state delay theory, $d_v = \frac{Sr^2}{2c(S-q)}$, the average delay at expressway is

$$d_k = d_c + \frac{Sr^2}{2c(S-q)} = d_c + \frac{Sd_c^2}{2(d_c+t_f)(S-q_k)} \quad (4)$$

Therein, q is the vehicle arrival rate of a certain section.

- q_k is average traffic volume at expressway. t_f is the dissipating time of the queue length on the exit ramp, and the unit is s. $t_f = \frac{q_k d_c}{(S-q_k)}$.
- r is virtual time of red light in a circle, and its value is d_c that is the average delay of the off ramp. The unit is s.
- g is virtual time of green light in a circle, and its value is the dissipating time of the queue length on the exit ramp.
- c is the virtual time of one cycle, is the sum of the virtual time of red light and green light. The unit is s.

If queue vehicles at expressway main road do not ascend to upstream on ramp, the traffic flow of on ramp can enter into expressway main road with critical gap, and then the delay of on ramp can be obtained by critical gap theory. Expressway section traffic volume is assumed as negative exponential distribution, and then probability of time headway greater than 6 seconds is $P(h \geq 6) = e^{\frac{-q_k \times 6}{3600}}$.
The number of space headway that expressway can provide to in-ramp in one hour $N = q_k \times e^{\frac{-q_k \times 6}{3600}}$.
Average waiting time of traffic flow on on ramp entering into expressway is $d_r{}' = \frac{3600}{N}$.
However, if queue length on expressway ascend to on ramp and it leads to on ramp traffic can't enter into the main road, average delay of traffic flow on on ramp entering into expressway includes the delay caused by queue dissipation time on off ramp is d_c, the delay on side road caused by traffic flow dissipation of expressway main road is d_k, and the delay caused by vehicles on on ramp entering into the main road is d_{r1}. And then $d_r = d_k + d_{r1}$.
Therein, d_{r1} is calculated by steady state theory, red time is d_k, green time is dissipation time of expressway queue length, that is $\frac{q_k \times (d_c + t_f)}{(S-q_k)}$, and then

$$d_{r1} = \frac{Sr^2}{2c(S-q)} = \frac{Sd_k{}^2}{2\left(d_k + \frac{q_k \times (d_c + t_f)}{(S-q_k)}\right)(S-q_r)} \quad (6)$$

Therein, q_r is the average traffic flow of the on ramp. The unit is veh/s.

On ramp delay is

$$d_r = \begin{cases} \frac{3600}{q_k \times e^{\frac{-q_k \times 6}{3600}}} + d_h \; q_k \times (d_c + t_f) \times 7 + 20 \leq L_s \\ d_k + d_{r1} \; q_k \times (d_c + t_f) \times 7 + 20 > L_s \end{cases} \quad (7)$$

Therein,

- d_r is the average vehicle delays at the on ramp on the side road. The unit is veh/s.
- L_s is the distance between off ramp and on ramp on expressway.

The queue on on ramp will produce a certain delay to side road traffic of on ramp upstream. But considering that a part of side road traffic flow has been diverted into expressway main road at this time, the rest has existed queue because of signal control at the intersection downstream. Once queue vehicles on the on ramp dissipate, side road congestion will be relieved in a certain degree. Therefore, the delay of side road caused by expressway is considered that expressway behaves a lane to downstream intersection, the delay of expressway is computed , and delay of side road is mainly caused by the downstream intersection , is mutual independent by other lanes.
Therefore, under the first into then out condition, average delay of traffic flow is,

$$d_z = \frac{q_c d_c + q_k d_k + q_r d_r}{q_c + q_k + q_r} \quad (8)$$

Therein, d_z is the average delays in all vehicles on expressways and side roads under the on-off ramp condition. The unit is s.

From the above model, it can be seen that expressway traffic operation state is closely related to signal circle at the downstream intersection, off ramp number, traffic volume of side road, traffic volume of expressway main road, distance between on and off ramp and traffic volume of on ramp. Among them, traffic volume parameters are a class of traffic parameters, parameter greatly affected by time is signal timing at the downstream intersection, parameter greatly affected by space is the distance between on and off ramp.

2.2. Off-on ramp group

From Figure 1(b), it can be seen that once on ramp queuing occurs, and ascend to the side road, it affects the normal operation of the side road traffic and produce corresponding queue and delay. At the same time, if the queue length of side road reaches off ramp, it will cause that vehicles of off ramp can not pull out of the expressway, and then the queue length of off ramp will be affected and expressway traffic will be disturbed.

Based on the analysis, using critical gap theory and queuing theory, the expressway delay model and side road delay model can be established based on the queue length of on and off ramp.

According to the delay model when traffic flow do not ascend to on ramp under the first into then out condition, the delay of on ramp under the first out then into condition can be calculated by

$$d_r = \frac{3600}{N} + d_h = \frac{3600}{q_k \times e^{\frac{-q_k \times 6}{3600}}} + d_h \tag{9}$$

Then queue length on on ramp is the number of vehicles entering into on ramp during the period of time between two adjacent critical gaps. The equation is $L_r = d_r \times \frac{3600}{N}$.

on time on on ramp is $t_r = \frac{q_r \times \frac{3600}{N}}{S}$.

Both queue length caused by vehicles of on ramp and queue dissipation time influence on traffic flow on side road and subsequent traffic greatly. During this time there is at least one lane traffic affected, and the influencing vehicles can pass this bottleneck after queue dissipation.

The paper assumes that the approach of off ramp is one lane, and only one lane on side road is disturbed, however, other lanes are not affected. Therefore, delay of the lane on side road can be calculated by the delay model of signal intersection. Red time is $r=d_r+t_r$, and green time is $g = \frac{q_f(d_r+t_r)}{S}$.

Average delay of the single lane on side road is

$$d_f = \frac{Sr^2}{2c(S-q)} = \frac{S(d_r+t_r)^2}{2\left(d_r+t_r+\frac{q_f(d_r+t_r)}{S}\right)(S-q_f)} \tag{10}$$

If queue length of side road is too long and queue traffic ascends to off ramp, traffic flow of off ramp entering into side road will be affected. Therefore, a certain delay can exist on off ramp and the delay is closely related to dissipation time on side road. If queue length on side road dissipates completely and there are critical gaps, traffic flow on on ramp can enter into side road.

Delay on off ramp is shown in equation 11.

Similarly, there is a certain delay on expressway main road because of traffic flow on off ramp. The delay model is closely linked with queue dissipation time of off ramp, and average delay on the right lane of expressway main road is,

$$d_k = d_c + \frac{Sd_c^2}{2(d_c+\frac{q_k d_c}{S})(S-q_k)} \tag{12}$$

Therefore, under the first out then into condition, average delay of traffic flow is,

$$d_z = \frac{q_c d_c + q_f d_f + q_k d_k + q_r d_r}{q_c + q_f + q_k + q_r} \tag{13}$$

From the above model, it can be seen that expressway traffic operation state is closely related to traffic volume of in ram, traffic volume of side road, traffic volume of expressway main road, and distance between on and off ramp. Among them, traffic volume parameters are a class of traffic parameters, there is no parameter affected by time, parameter greatly affected by space is the distance between on and off ramp.

3. Safety Analysis based on traffic operation state Expressway

According to the above research, we establish the relationship between the related traffic volume, the distance between on and off ramp and the total delay, and the total delay and the expressway safety are closely related.

Under the small traffic volume condition, vehicle speed is the key factor affecting road safety, however, under the large traffic volume condition, traffic volume and congestion degree are the important factors affecting road safety.

$$d_c = \begin{cases} d_r + t_r + \frac{q_f(d_r+t_r)}{S} + \frac{3600}{q_f \times e^{\frac{-q_f \times 6}{3600}}} + d_h\, q_f \times \left(d_r + t_r + \frac{q_f(d_r+t_r)}{S}\right) \times 7 + 20 > L_f \\ \frac{3600}{q_f \times e^{\frac{-q_f \times 6}{3600}}} + d_h q_f \times \left(d_r + t_r + \frac{q_f(d_r+t_r)}{S}\right) \times 7 + 20 \le L_f \end{cases} \tag{11}$$

When traffic volume is large, the probability of conflictions between vehicles becomes large. In the process of confluence and shunting, traffic accidents are easy to taken, such as cut rub, because of different driving technology, and road safety reduces sharply.

Compared with the urban road there is not signal intersection on expressway, but compared with the highway, the number of on and off ramp is far more than that of the highway. Hence as a relatively special type of road, the safety of expressway is mainly reflected under the large traffic volume. When the traffic volume is small, the safety of expressway is similar with that of highway, however, when the traffic volume is large, the safety of expressway is different from that of urban road because there is no signal intersection to ensure safety on expressway. Therefore, the paper mainly studies the effect of traffic congestion degree on expressway safety.

According to traffic control principle, when traffic flow ratio is less than 0.7 traffic is generally considered to be smooth, when traffic flow ratio is between 0.7 and 0.8, traffic flow begins to be in congested state, and when traffic flow ratio is more than 0.8, traffic flow is in saturation state and vehicles can not pass. It can be seen that with traffic volume increasing, road safety is changing to poorer, especially when traffic flow ratio is between 0.7 and 0.8, acceleration and deceleration behavior of a certain driver perhaps will lead to traffic accident and even the whole expressway network safety decreased.

Likewise, when on and off ramps delay increasing sharply expressway safety is reduced gradually, while when the delay is below the delay of smooth traffic flow expressway safety is relatively high. According to this principle, the paper established the model of on and off ramp delay and expressway safety.

The safety in this paper refers to probability of vehicles collision occurred. If collision probability is great the safety is poor, and if collision probability is small the safety is high. The paper argued that probability of vehicles collision occurred is mainly affected by delay and traffic flow ratio. Therefore, with delay and traffic flow ratio increasing collision probability is increasing. If traffic flow ratio is small and the congestion of on and off ramp is heavy, collision probability will be increasing, and if traffic flow ratio is great and the delay of on and off ramp is small, collision probability will be increasing too. Therefore, the safety model of accident occurrence probability based on the ramp is established. The principle is when a certain traffic flow delay occurs, the traffic flow is in congested and low speed, and it is similar to increasing some vehicles. That is to say, based on the original traffic flow, the delay is equivalent to increase some vehicles. Thus traffic volume is increased and safety is decreased. That is,

$$P(l) = \frac{q_k(1 + d_z/d_{max})}{C} = \begin{cases} \frac{q_k\left(1+\frac{q_c d_c + q_k d_k + q_r d_r}{\sum q \times d_{max}}\right)}{C} & l_{rc} \le l_{cmin} \\ \frac{q_k\left(1+\frac{q_c d_c + q_f d_f + q_k d_k + q_r d_r}{\sum q \times d_{max}}\right)}{C} & l_{cr} \le l_{rmin} \end{cases} \quad (14)$$

Therein,

- $P(l)$ is the accidents occurrence probability based on on and off ramp.
- d_z is the average delay produced by the weaving section , and the unit is s. Under the on-off ramp condition, d_z is $q_c d_c + q_k d_k + q_r d_r$, which is calculated by the delays and traffic volume of off ramp, expressway main road and on ramp. Under the off-on ramp condition, d_z is $q_c d_c + q_f d_f + q_k d_k + q_r d_r$, which is calculated by the delays and traffic volume of off ramp, expressway main road , on ramp and side road.
- $\sum q$ is the sum of all traffic flows delayed by weaving section, and the unit is *veh/h*. Under the on-off ramp condition, $\sum q$ is the total traffic flow of on-off ramp and expressway main road. Under the off-on ramp condition, $\sum q$ is the total traffic flow of on-off ramp, expressway main road and side road.
- d_{max} is the maximum delay caused by traffic flow in expressways under congestion condition, and the unit is *s*. The maximum delay value is proportional to congestion time. If the arrival traffic volume is increasing, congestion continues to deteriorate, and traffic delays continue to increase. If the arrival traffic flow is less than the departure flow, the traffic congestion begins to dissipate and the delays are decreasing gradually. Thus in order to reduce the effect of congestion time on the maximum delay, the maximum delay is calculated in the condition the queues begin to

dissipate when arrival traffic volume accumulated to the maximum leaving vehicle number. Do not considering increasing congestion as the number of arrival vehicles continues to increase in the condition of over saturation. In this assumption, it is possible to determine that the maximum delay of traffic flow is the cycle time of virtual cycle. Under the on-off ramp condition, the maximum delay $d_{max} = d_k + d_c + \frac{q_k(d_c+t_f)}{S-q_k}$. Under the off-on ramp condition, the maximum delay is $d_{max} = c = r+g = d_r+t_r+\frac{q_f(d_r+t_r)}{S}$.

- l_{rc} is the distance from the upstream on-ramp to the downstream off-ramp according to the traffic driving direction, and the unit is m.
- l_{cr} is the distance from the upstream off-ramp to the downstream on-ramp according to the traffic driving direction, and the unit is m.
- l_{cmin} is the minimum distance from the upstream on-ramp to the downstream off-ramp when the traffic flow at upstream on-ramp cannot affect the traffic flow at downstream off-ramp according to the traffic driving direction, the unit is *m*. $l_{rc} \leq l_{cmin}$ represents that the distance from the upstream on-ramp to the downstream off-ramp is less than the minimum distance and the traffic flow at on-ramp will affect the vehicles at off-ramp.
- l_{rmin} is the minimum distance from the upstream off-ramp to the downstream on-ramp when the traffic flow at upstream off-ramp cannot affect the traffic flow at downstream on-ramp according to the traffic driving direction, the unit is *m*. $l_{cr} \leq l_{rmin}$ represents that the distance from the upstream off-ramp to the downstream on-ramp is less than the minimum distance and the traffic flow at off-ramp will affect the vehicles at on-ramp. *C* is traffic capacity, and the unit is *veh/h*.

4. Simulation verification

As reflected factor of road traffic accident expressway safety random is obvious, simulation is not easy to achieve. Therefore, the paper assumed parameters on expressway, calculated different type of delay models at on and off ramp based on the above models, and the safety index on expressway of different types is obtained to judge whether the expressway situation assumed is consistent with the ultimate safety index, and then judge whether the model is feasible.

The paper assumed that on-off ramp group, traffic volume of single line on expressway main road is 900 *veh/h*, traffic volume of single line on side road is 600 *veh/h*, traffic volume of on ramp is 500 *veh/h*, traffic volume of off ramp is 400 *veh/h*, the distance between signal intersection of off ramp downstream and off ramp is over 300 meters and downstream signal intersection does not affect the traffic flow of the off ramp greatly. The distance between on and off ramp is only 100 meters, and the traffic flow is high. This paper would computer the safety of expressway in this condition to validate the accuracy of the new safety model.

Firstly delay on off ramp is calculated by

$$d_c = \frac{q_c+\overline{q_f}}{\overline{q_f}(\overline{q_f}-q_c)} = 30s \quad (15)$$

$$t_f = \frac{L_q}{\mu} = 30s \quad (16)$$

And delay on expressway is determined based on delay on off ramp, that is

$$d_k = d_c + \frac{Sd_c^2}{2(d_c+t_f)(S-q_k)} = 45s \quad (17)$$

$$d_{max} = d_k + \frac{q_k(d_c+t_f)}{S-q_k} = 99s \quad (18)$$

According to delay and dissipation time on off ramp, queue length on expressway is calculated by

$$q_k \times (d_c + t_f) \times 7 = 105s \quad (19)$$

That is more than 100 meters the distance between on and off ramp. It indicates that queue length on expressway ascends to the upstream on ramp, then

$$d_{r1} = \frac{S(d_k+d_c)^2}{2\left(d_k+d_c+\frac{q_k\times(d_c+t_f)}{(S-q_k)}\right)(S-q_r)} = 10.86s \quad (20)$$

$$d_r = d_k + d_{r1} = 55.86s \quad (21)$$

So the average delay of all traffic flow is,

$$d_z = \frac{q_c d_c+q_k d_k+q_r d_r}{q_c+q_k+q_r} = 44.68s \quad (22)$$

Finally, the accidents occurrence probability of expressway is calculated as 0.73.

$$P(l) = \frac{q_k(1+d_z/d_{max})}{C} = 0.73 \quad (23)$$

It can be seen that owing to the delays on the side road is not counted, the average delay of the whole traffic flow is not greatly increased. But supposed traffic flow ratio on expressway is 0.5 that is amount to the accidents occurrence probability of 50%, and at the on and off ramp, the accidents occurrence probability of expressway is 73%, which shows the degree of danger is increased by 46%, and the safety of expressway is decreased compared with the safety in the smooth flow condition.

5. Conclusions

As the very important parts of the expressway system, on and off ramps have a great effect on the operation effect of expressways. Once congestion occurs at on and off ramps, it will directly affect the safety of the expressway and vehicles. Therefore, this paper aims to analyze the influence on expressway safety in aspect of on and off ramps. Firstly, we analyzes the location design of on ramps, the length and the influences on the expressway running state of off ramp downstream traffic state. In addition, we establishes the expressway operation models to analyze delay of different matches of on and off ramps, which include off ramp, on ramp, expressway and side road delay model under the influence of the distance between on and off ramps and traffic volumes. Meanwhile, the connection between the traffic operation state and traffic safety is analyzed, and the expressway safety model based on congestion degree is established. Through simulation verification, the impact of on and off ramp on the security of the expressway is analyzed. Finally from the results it can be concluded that the delay caused by traffic of the on and off ramps can decrease the safety of expressway.

References

[1] AL-KAISY, A. F., HALL, F. L. & REISMAN, E. S. 2002. Developing passenger car equivalents for heavy vehicles on freeways during queue discharge flow. *Transportation Research Part A: Policy and Practice,* 36**,** 725-742.

[2] BARED, J., GIERING, G. L. & WARREN, D. L. 1999. Safety evaluation of acceleration and deceleration lane lengths. *Institute of Transportation Engineers. ITE Journal,* 69**,** 50.

[3] CHEN, H., LIU, P., LU, J. J. & BEHZADI, B. 2009. Evaluating the safety impacts of the number and arrangement of lanes on freeway exit ramps. *Accident Analysis & Prevention,* 41**,** 543-551.

[4] GAO, J. & GUO, X. 2006. Research on the Space Optimizing Design Between Ramp of Urban Viaduct Road and Intersection [J]. *Road Traffic & Safety,* 10**,** 002.

[5] HANDKE, N. 2010. Possibilities and duties of ITS for large events. *Archives of Transport System Telematics,* 3**,** 19-22.

[6] JAYAKRISHNAN, R., TSAI, W. K. & CHEN, A. 1995. A dynamic traffic assignment model with traffic-flow relationships. *Transportation Research Part C: Emerging Technologies,* 3**,** 51-72.

[7] KAROŃ, G. & ŻOCHOWSKA, R. 2015. Modelling of expected traffic smoothness in urban transportation systems for ITS solutions. *Archives of Transport,* 33(1), 33-45.

[8] KOJIMA, M., KAWASHIMA, H., SUGIURA, T. & Ohme, A. The analysis of vehicle behavior in the weaving section on the highway using a micro-simulator. Vehicle Navigation and Information Systems Conference, 1995. Proceedings. In conjunction with the Pacific Rim TransTech Conference. 6th International VNIS.'A Ride into the Future', 1995. IEEE, 292-298.

[9] LEE, C. & ABDEL-ATY, M. 2006. Temporal variations in traffic flow and ramp-related crash risk. *Applications of Advanced Technology in Transportation.*

[10] LEISCH, J. E. 1959. Spacing of Interchanges on Freeways in Urban Areas. *Transactions of the American Society of Civil Engineers,* 126**,** 604-616.

[11] LIU, P., CHEN, H., LU, J. J. & CAO, B. 2009. How lane arrangements on freeway mainlines and ramps affect safety of freeways with closely spaced entrance and exit ramps. *Journal of Transportation Engineering,* 136**,** 614-622.

[12] MUNOZ, J. C. & DAGANZO, C. F. 2002. The bottleneck mechanism of a freeway diverge. *Transportation Research Part A: Policy and Practice,* 36**,** 483-505.

[13] PULUGURTHA, S. S. & BHATT, J. 2010. Evaluating the role of weaving section characteristics and traffic on crashes in weaving areas. *Traffic injury prevention,* 11**,** 104-113.

[14] QU, X., YANG, Y., LIU, Z., JIN, S. & WENG, J. 2014. Potential crash risks of expressway on-ramps and off-ramps: a case study in Beijing, China. *Safety science*, 70**,** 58-62.
[15] WANG, L., SHI, Q. & ABDEL-ATY, M. 2015. Predicting crashes on expressway ramps with real-time traffic and weather data. *Transportation Research Record: Journal of the Transportation Research Board***,** 32-38.

ARCHIVES OF TRANSPORT www.czasopisma.pan.pl PAN www.journals.pan.pl ***ISSN (print):*** *0866-9546*
Volume 43, Issue 3, 2017 ***e-ISSN (online):*** *2300-8830*
DOI: *10.5604/01.3001.0010.4227*

FORECASTING TRAVEL TIME RELIABILITY IN URBAN ROAD TRANSPORT

Mattias Juhász[1], Tamás Mátrai[2], Csaba Koren[3]

[1, 3] Department of Transport Infrastructure, Széchenyi István University, Győr, Hungary

[2] Department of Transport Technology and Economics, Budapest University of Technology and Economics, Budapest, Hungary

[1]e-mail: mjuhasz@sze.hu
[2]e-mail: tamas.matrai@mail.bme.hu
[3]e-mail: koren@sze.hu

Abstract: *Assessment of travel time reliability as a fundamental factor in travel behaviour has become a very important aspect in both transport modelling and economic appraisal. Improved reliability could provide a significant economic benefit if it is adequately calculated in cost-benefit analyses for which the theoretical background has already been set. However, methods to forecast travel time reliability as well as travel behaviour models including its effects are rather scarce and there is a need for development in this field. Another important aspect could be the influencing factor of reliability in travel demand management and related policy-making. Therefore, this paper intends to further analyse reliability focusing exclusively on urban road transport based on automatic measurements of journey times and traffic volumes from a dataset of the city of Budapest. The main finding and the novelty of the study is a model which can forecast the standard deviation of travel times based on the volume-capacity ratio and the free-flow travel time. The paper also provides a real-life numerical experiment in which the proposed model has been compared with other, existing ones. It proves that besides existing mean-delay-based models, travel time reliability can be forecasted based on the volume-capacity ratio with an adequate accuracy.*

Key words: *travel time reliability, forecasting, urban road transport, appraisal, congestion.*

1. Introduction

Management and planning of urban transport systems is a complex task which demands a comprehensive approach in supporting decision-making. In order to make 'well-informed' decisions (e.g. choose the best alternatives in developing the system), it is indispensable to take into consideration every relevant aspect and effect of the given interventions. For this purpose, there are different policy and project assessment tools. A universal, widely accepted and long-standing tool is cost-benefit analysis (CBA) which is mainly assessing projects from an economic point of view. However, one can argue that the method itself has its own limitations and there are important effects which can be hardly monetized (or just quantified). Previous papers reviewed these limitations and challenges ahead for appraisal methods (Mátrai and Juhász, 2012; Mátrai, 2013). Major transport economists accept that there are new, innovative methods, but most of them believe that with methodological additions and proper quality of implementation CBA still allows the most prudent form of analyses to be carried out. (Laird et al., 2014; Vörös et al., 2015)

In recent years travel time reliability (TTR) is an increasingly important issue among transport experts (ITF, 2010). One of the leading international organizations – OECD – organized a roundtable in late 2015, where several transport economists provided their view on this issue (Kouwenhoven and Warffemius, 2016; Fosgerau, 2016). Travellers intend to optimize their daily activity chains (Esztergár-Kiss and Rózsa, 2015), which results in shorter travel times and they are also sensitive to the variability (predictability) of travel times as unreliability press users to use safety margins (buffer times) which cause an additional disutility (travel cost) beyond pure travel time. Therefore, TTR is a fundamental factor in understanding and modelling travel behaviour. Furthermore, improved reliability could be a significant economic benefit if it is calculated in CBAs. Several countries have recently decided to include TTR in their CBA guidelines and defined the monetary values (i.e.

value of reliability, VOR). However, methods to predict the impact of interventions on TTR (reliability forecast models) as well as travel behaviour models including TTR effects are also needed. These models are still rather scarce and there is no concord among professionals on which method should be used. (Eliasson, 2006; de Jong and Bliemer, 2015).

Another important aspect could be the influencing factor of reliability in travel demand management and related policy-making. On the one hand, in case of restrictive road projects (e.g. traffic calming projects) a decrease in reliability could mean an undesirable side effect (a loss for the society). On the other hand, in case of a public transport or non-motorized development, modal shift can have a positive effect on overall TTR. Moreover, reliability of travel times could also influence land-use decisions, so it should be a factor in land-use and transport interaction modelling as well (Juhász, 2014). TTR can be also important for analysis of cycling investments, since the mode shift from car to cycling is usually marginal. In absolute terms the number of cars decreases only with a small amount which provides nearly no impact on travel times, but might have a significant one on reliability.

Nowadays the pervasive development of info-communication technologies and intelligent transport systems (e.g. intelligent sensors) provides the opportunity to further analyse TTR and expand possibilities in forecasting. These investigations should be focused on urban regions for two important reasons: (1) more than 50% of the world population is living in an urban area and according to the general predictions this ratio is expected to further increase (United Nations, 2015); (2) congestion (and unpredictability) as a major transport issue are mainly concentrated in cities (ITF, 2010; Rao A.M. and Rao K.R., 2012). As well-established and widely accepted guidelines are missing on how to forecast TTR, there is a need for further analysis. This study focuses exclusively on urban road transport as it is presumed that the reliability issue is mostly significant in this setting, however it is likely to be relevant in other circumstances (such as for long-distance or public transport trips) as well (Eliasson, 2006; Spławińska, 2015; Horbachov et al., 2015). The database of road operators can be a platform to measure reliability as road authorities often collect data of traffic volumes and individual vehicle trips for different purposes (e.g. to provide information for road users on estimated travel times to a certain destination). From these dataset, the reliability of a given route can be characterized, if the traffic situation (e.g. saturation level) is also known.

Based on the aforementioned aspects and focusing on urban road transport, the objective of this paper is to:

- provide a brief review on TTR approaches;
- explore the relation between TTR and relevant traffic parameters based on the case study of Budapest, in which automatic travel time measurement of a traffic information system has been used;
- propose a methodology to forecast TTR and draft further research.

2. Review of TTR approaches in transport appraisal

The topic of TTR has been investigated by numerous studies. The history of the research is summarised by Taylor (2013). In this section a brief review is provided from the most relevant papers to describe the background of the topic and this research.

Travel times of road trips are usually not stable over time. Variations occur as a consequence of fluctuations in travel demand and road capacity. A part of these fluctuations is known to road users (e.g. regular, cyclical variations), while another part is not (irregular or random variations). This paper focuses on unexpected variations which can cause that road users arrive earlier or later than expected. The unreliability forces travellers to add buffer times to their trips in order to avoid being late. This is then an additional disutility (cost) to mean travel time. But in some cases standard buffers (‘head starts’) could be insufficient and lateness could also cause another disutility, while arriving too early can also have its unpleasantness. So unreliability could mean a cost for users as they might face additional travel times, lateness, waiting times and even they may need to reschedule their activities. All of these might be accompanied by bad feelings such as anxiety. (Dale et al., 1996 ; Bates et al., 2001; Peer et al., 2010; Taylor, 2013)

TTR is the level of unpredictable, day-to-day variation of travel times, which represents the temporal uncertainty experienced by road users during their trips and it is related to transport

network conditions in a complex way. In this interpretation, reliability is basically equivalent to the predictability of travel times and associated with the statistical concept of variability (Kouwenhoven and Warffemius, 2016). A high level of reliability means a low level of variability (uncertainty), which indicates that travel times are mostly predictable. However, reliability can also be approached from the aspect of expectations. In that regard, users are expecting (!) a certain level of service upon which they organize their activities and reliability is proportional to the ability of the transport system to fulfil this requirement. Travel time variability can be caused by special events (e.g. accidents) as well but in this paper the focus is purely on day-to-day variations which mainly arise in congested situations as several studies found that the main explanatory factor of travel time variability is mean travel time (i.e. the sum of free flow travel time and mean delay). In a severe congestion, this variation might be very significant, but if it is very severe, variability can also become a decreasing function of travel time as in a very congested situation travel times are mostly homogeneous (Eliasson, 2006; Mátrai 2012). By improving TTR, additional 'buffer' times, waiting times and the probability of lateness could be decreased. For certain projects, not including TTR in economic calculations, a significant benefit or loss may be disregarded. That is why the economic impacts of TTR changes are more and more about to appear in CBAs. Travel time-related benefits are traditionally measured as the improvement of journey times. With incorporating reliability, those time benefits need to be split into (conventional) travel time savings and savings on TTR. Different studies proved that reliability could have a very significant effect in CBAs. Previous papers (e.g. Mátrai and Juhász 2012) pointed out that including reliability benefits in a public transport investment could add to 8-15% to the economic benefits, while others such as Eliasson (2006) or Kouwenhoven and Warffemius (2016) found that in road investments it could also add 10-60% to the benefits. In order to calculate the economic impact of TTR in a given project the following steps are needed based on the papers of Kouwenhoven and Warffemius (2016) and Fosgerau (2016):

1) determination of a monetary value of reliability (VOR – the cost to travellers per unit of travel time variability),
2) measurement and prediction of the level of TTR (the quantity of travel time variability),
3) incorporating the reaction of users to reliability in travel behavioural models (e.g. in route choice models by including the cost of variability into the generalised travel cost function).

In order to go through the aforementioned steps, first and foremost a unit of measurement needs to be defined for TTR. Approaching the topic from this operational (measurement) aspect, there are two main groups of definitions for reliability. The first one is the 'mean-dispersion' model which defines a measure of dispersion of travel time distribution (standard deviation, variance range, percentiles etc.). In this model the standard utility function contains the travel cost, the travel time and the dispersion of travel time. Value of time (VOT) can be defined as the marginal rate of substitution between travel time and travel cost, while VOR is the rate of substitution between reliability and travel cost. The latter represents the monetary value travellers place on improving the predictability of travel times (i.e. reducing the travel time variability). The second group is the 'scheduling delay' model, in which the scheduling consequences of TTR are measured by the expectations of arriving or departing earlier or later then the preferred time. (Dale et al., 1996; Eliasson, 2006; Fosgerau and Hjort, 2008; Fosgerau et al., 2008; TRB, 2011; de Jong and Bliemer, 2015).

Having reviewed the literature on TTR approaches, it seems that at this stage (!) standard deviation of travel time is the most appropriate way to quantify the variability of travel times within the mean-dispersion model. It seems the era of scheduling models is still to come as there are only a limited number of practical researches on this field. Furthermore, the data on the preferred arrival time of the users is very limited which is a prerequisite of using the scheduling model. Then the mean-dispersion model can be applied and the only question is how to measure the dispersion within. Numerous studies pointed out that travel times are not normally distributed and there is an evidence of skewness to the upper tail (Taylor 2013; Susilawati et al. 2013; de Jong and Bliemer 2015). Due to this fact some papers like Eliasson (2006) or de Jong and Bliemer (2015) suggested to not use the standard deviation as a measure of dispersion because it is affected by this skewness, as it is basically (and

more appropriately) used for symmetrically distributed variables. Therefore, these papers suggested to use the difference of specific quantile values or difference between quantiles and the mean travel time. However, Fosgerau (2016) showed that the standard deviation is proportional to the other measures of dispersion such as differences of specific quantiles. Moreover, he also stated that theoretically standard deviation is more appropriate for commuters with inflexible working times which are more general among travellers in an urban peak hour. Standard deviation has several advantages: (1) it is easy to estimate, (2) it is easy to include in a transport model and in a CBA, and (3) it is the most common TTR measurement in practice. However, there are further arguments against as there are difficulties in calculations over routes as it is not an additive formula (only variances of links can be summarized). Another mode of visualisation of travel time reliabilities and changes is the rubber sheet method used on travel time maps (Ficzere et al., 2014; Kouwenhoven and Warffemiu,s 2016). Based on the consideration of Fosgerau and due to the limited data on preferences required for scheduling models, in this research the standard deviation is used as a proxy for TTR.

The conceptual models of the valuation of travel time variability (basic model, step model, slope model) are summarised by Fosgerau (2016), while a study from de Jong and Bliemer (2015) provides a comprehensive review on deriving VOR and on TTR forecast models. VOR is mostly expressed as the product of VOT and a reliability ratio. Based on this review, reliability ratios are mostly in the range of 0.4 and 1.1 for passenger transport, while it is usually a bit lower (between 0.1 and 0.4) for freight transport. In terms of TTR forecast models the study mentions seven national methods from which five is summarized here by Table 1, in which 'D' is the distance; 'MD' is the mean delay, 's' and 's_0' is the maximum and minimum of standard deviation of travel times respectively; 'std' is the standard deviation of travel times; 't' is the (mean) travel time; 't_0' is the free-flow travel time, 'v' is the speed; 'F/C' is the traffic volume (flow) – capacity ratio, while 'a', 'b', 'c' and 'd' are constant parameters.

Most of these models calculate the standard deviation of travel times based on the estimated mean delay (ratio of mean travel time and free-flow travel time) as an indicator of congestion. All of these models are quite useful to measure TTR impacts (standard deviation), but the problem is that TTR cannot be incorporated to a standard assignment model in a way that it is depending on the mean travel time as it is calculated within the process. So the main issue is the interdependence between these values. Furthermore, it is important to note that some interventions could have different impacts on mean travel time and TTR, which also suggests to forecast the standard deviation independently of the mean delay. The model from New Zealand is an exception as it calculates the standard deviation based on the extreme values and the F/C ratio. This model avoids the issue of TTR and mean delay interdependence (the 'endogeneity' issue), but it is also impossible to incorporate TTR to assignment models this way as the model uses the extreme values of standard deviation, which cannot be properly estimated beforehand for future years.

3. Forecasting reliability in urban road networks

Based on the aforementioned aspects of existing methods, this study aimed at developing a model to forecast the standard deviation of travel times based purely on the F/C ratio and the free-flow travel time. This approach is basically parallel to developing volume-delay functions (VDFs) which describe the

Table 1. Summary of existing TTR forecast models based on de Jong and Bliemer (2015)

#	Name of model (nation)	Expression
1	Arup 2003 (UK)	$std = 0.148\ MD^{0.781}\ D^{-0.285}\ t$
2	NZTA 2010 (New Zealand)	$std = s_0 + (s-s_0) / (1 + \exp[a\ (F/C - 1)])$
3	Kouwenhoven et al. 2005 and Kouwenhoven, Warffemius 2016 (The Netherlands)	$std = a + b\ MD + c\ \ln(MD+1) + d\ D$
4	Eliasson 2006 (Sweden)	$std = t \exp(a + b\ (MD-1) + c\ (MD-1)^3)$
5	Geistefeldt et al. 2014 (Germany)	$std = a\ MD^b$

mean-delay based on the saturation level and the free-flow travel time of a link. In this way it would be easy to calculate an index of TTR (standard deviation) within a standard macroscopic transport model and also to incorporate reliability into the choices of travellers. In this paper it was also intended to develop this new method and compare the results with those of the existing models.

3.1. The concept

On a macro transport modelling level, the widely-used VDFs are describing the expected values of travel time (or mean delays). These functions provide fairly good estimations and a better way of estimation is still to be discovered. However, existing urban transport models usually does not calculate the standard deviations of these expected travel times (so that TTR). Therefore, and based on the review of existing TTR forecast models, the essential concept was to develop a similar function to VDFs to determine the relationship of traffic volume and TTR. With such a function TTR could be forecasted for 'do-nothing' (reference) and 'do-something' (project) cases during a transport modelling procedure. As reliability is generally affected by the level of congestion, i.e. the F/C ratio, a universal parameter of traffic state has been chosen as an explanatory variable for two reasons: (1) it is easy to calculate it in a transport model; (2) it can purely represent the level of congestion without using other estimated and interdependent values such as mean delay. To this end and based on Taylor (2013) a longitudinal data collection was needed in which trip times and saturation levels for given (preferably longer) urban routes were simultaneously measured. In case of the latter, counting the traffic volumes is enough as road capacities are known from design standards. Ultimately, assignment models and economic assessments can use the estimated values of standard deviations.

3.2. Data

In case of the city of Budapest (Hungary) it became possible to carry out the aforementioned experiment due to the so-called 'Easyway' project in which a traffic information system was implemented. A previous paper (Juhász et al., 2016) analysed the speed-flow relationship on urban roads which used the same data, therefore the description of it is based on that paper. On the inner section of M1-M7 motorway and main road No. 6 automatic number plate recognition cameras and variable message signs were installed in 2012 in order to inform the inbound traffic on the real-time average access time of the Danube bridges. Fortunately, the affected area is mostly covered with traffic-counting detectors which made it possible to measure traffic volumes on the road network. Figure 1 shows the measurement area.

A dataset from April 2014 was selected in order to carry out this research. The total number of measurements for the whole month is 525,000 (i.e. those trips for which it was possible to register both the travel time and the related traffic volumes). The automatic travel time measurement procedure classified the data into 6 and 15-minute time intervals for peak (from 4 a.m. to 5 p.m.) and off-peak periods respectively. Traffic volumes were registered by detectors in time intervals of 4, 8 and 10 minutes for peak (from 4 a.m. to 10 a.m. and from 12 a.m. to 6 p.m.), intermediate (from 10 a.m. to 12 a.m.) and off-peak (from 6 p.m. to 4 a.m.) periods respectively.

As travel time values were automatically rounded to minutes, whole routes were analysed because these rounded values are not characterising shorter road sections adequately. That was a severe limitation this research needed to face. Therefore, it was also needed to calculate route-level F/C values based on sectional ones. However, analysing whole routes has the advantage, that the results are easily comparable with drivers' expectations, as they think on the route level rather than on short section level.

3.3. The methodology

As transport modelling usually tries to represent a common, average setting of the transport system, its conclusions are mostly limited to generalized statements. While it would certainly worth to analyse the data of diverse time periods such as seasons or specific days, however, this study needs to follow the underlying generalization of transport modelling. Due to this consideration the dataset was filtered and days from Friday to Monday have been excluded.

As a first step the statistical solidity was checked for each measurement (for both the travel times and the traffic volumes). Due to failures or obviously wrong measurements some traffic counting locations were excluded from the analysis.

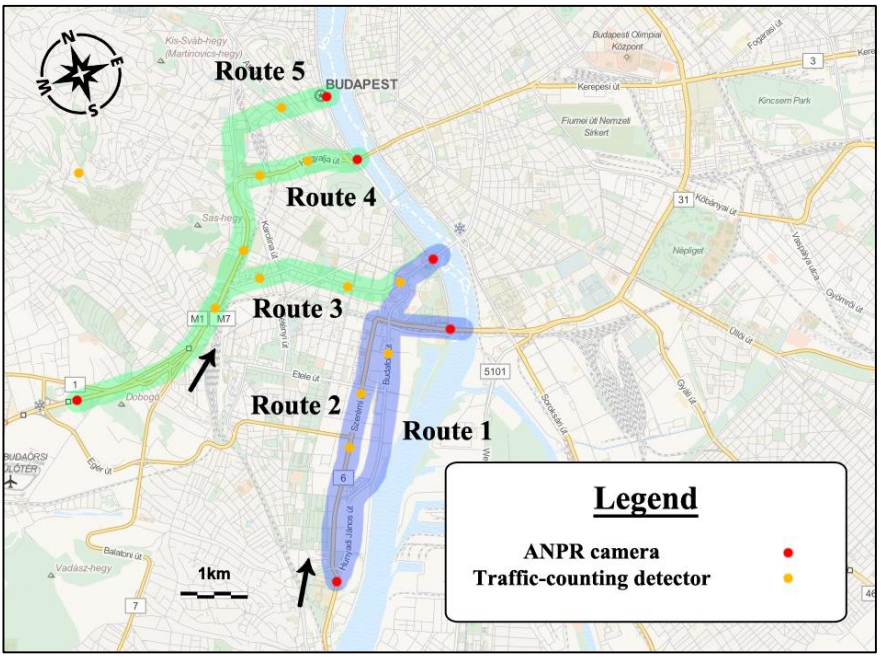


Fig. 1. The map of the measurement area (based on Juhász et al. 2016)

Extreme events were also excluded from the dataset based on Kouwenhoven and Warffemius (2016) in order to focus purely on day-to-day variations and to maintain consistency with underlying methods, as speed-flow curves and VOR stated preference surveys also exclude these extremes. Following the suggestions of the study, a boundary of exclusion was set to three times the raw standard deviation of travel times. As a consequence of filtering the extreme events (1,750 observations in total) a 3% decrease in the mean travel time and 15.4% in the standard deviation have been observed.

This study presents its results based on the data of route no. 4 as it had the most reliable dataset. The results were fairly similar on the other routes, but some lack of data and slight errors affected them. In case of route no. 4 around 80,000 measurements were available throughout the workdays that were involved in the analysis. The data coverage is shown by Figure 2. One can note that at least 100 measurements can be found in each saturation group of 5% and also in each 'hour of the day' group (the latter indicates the hour in which the measured car passed the starting point of the route) - Juhász et al. (2016).

Route no. 4 is a major – transit – route which is about 6.1 km long, starts at the end of a motorway and ends in the city centre. The selected route consists of different major road types. It means that minor and residential roads are not included but this should not be a problem as TTR is basically relevant on major urban roads. Speed limits are varying throughout the route (100-70-50 km/h as someone approach the city centre). In terms of intersections, there are six locations with traffic lights in a 24/7 mode and four pedestrian crossings without any signalization. Traffic volume is around 38,000 vehicles per day per direction on an average, from which around 20% is transit traffic. The morning peak is stronger, therefore the inbound direction was analysed in this work (see Figure 2) - Juhász et al. (2016)

In this study – contrary to other ones – the travel time dataset was not divided into specific (e.g. 15-minute) time intervals. Instead, F/C groups were created to calculate mean travel times and standard deviation as a relationship was sought between F/C values and standard deviation of travel times.

One of the major issues of this research was the difference between observed and modelled saturation ratios.

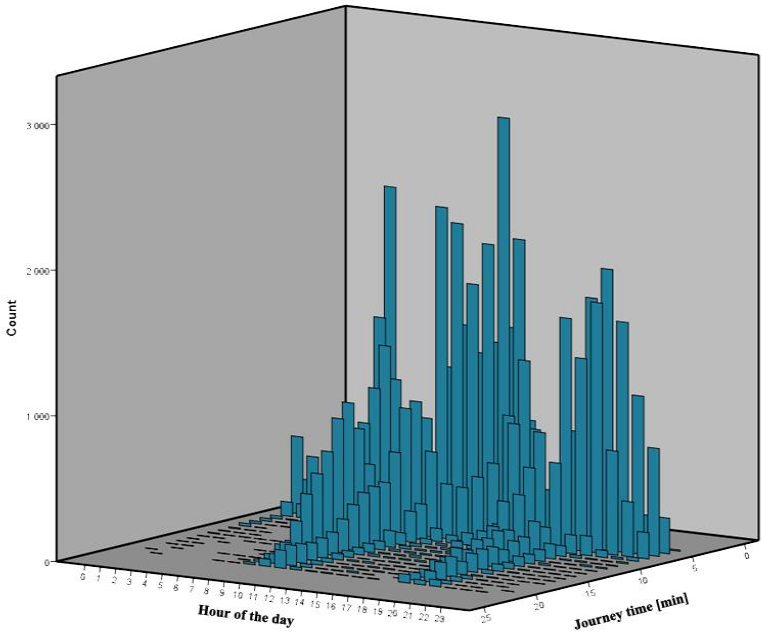


Fig. 2. The number of measurements according to the time of the day and journey times for route no. 4

Observed ones are calculated based on the actual traffic volume, but modelled values are in connection with travel demand, which means the number of users that are intended to use the road. The whole problem can be well-illustrated by the difference between two diagrams: the speed-flow diagram (the so-called fundamental diagram on the basis of Greenshields (1935) and Stamos et al. (2015)) and the standard VDF applied in transport modelling (see Ortúzar and Willumsen, 2011). In order to give an example, take a measured F/C value of 0.7 which can mean 0.7 in modelling if there is no congestion (labelled as 'normal state' and illustrated by point A in Figure 3) and a value above 1 if there is congestion (labelled as 'congested state' which is illustrated by point B). One can note that these traffic states are referred by different names in the literature: 'ordinary congestion' and 'hyper-congestion' are also in use respectively.

The task was to: (1) distinguish normal (not congested) states from congested states on the speed-flow curve; (2) find the proper F/C value in a modelling sense which can adequately represent a given congested state (point B' in the figure). In order to accomplish, first and foremost the validity of the speed-flow relationship for urban routes should be clarified. It was done in another part of this research and the results can be found in another paper (see Juhász et al., 2016). The applied method and relevant consequences are summarised in the following paragraphs.

'Normal' and 'congested' traffic states were distinguished with a method that analyse the dataset in a time sequence (going through each time step of the measurement). The classification method was defined based on the theoretical shape of the speed-flow curve (assuming that it is valid for this case based on Woollett et al. (2015) and Vasvári (2015). The theory suggests that a traffic state should be 'normal' if the F/C ratio and the mean travel time are changing in the same direction compared to the previous time step. And all other states should be labelled as 'congested' ones. Note that speed values of the fundamental diagram can be easily converted into journey times as the length of the route is given. However, within the literature congestion or congested states are often defined based on absolute or relative increases in travel times (e.g. in Eliasson, 2006).

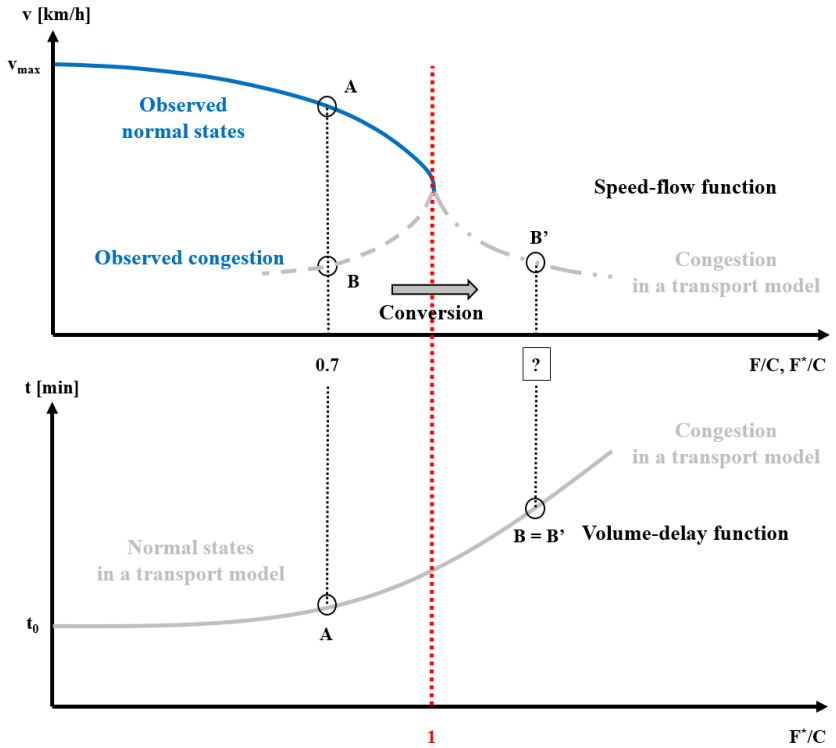


Fig. 3. The connection of the speed-flow relationship and the volume-delay function

In this work congested travel times cannot be analysed in a framework in which the state of congestion is defined on the magnitude of travel time which is a dependent variable. Therefore, this research defines 'normal' and 'congested' states based purely on the sign of the change. In this way the classifying model is not overdetermined compared to the previously mentioned methods.

As it was needed to analyse longer routes it was a difficult issue how to calculate the route-level F/C ratio in a given time interval. Each traffic counting stations characterise a shorter route section and if congestion starts to evolve in a section it needs time to spread to other sections. It is similar to the well-known wave propagation phenomenon from traffic flow theory (see Daganzo, 2007). Along the route three sections have been distinguished and characterised by traffic counting detector(s). Road capacity values were calculated based on the location of detectors. Having tried different methods to characterise the saturation level of the route, eventually the maximum of sectional F/C ratios were used. There were two reasons for that: (1) differences between sections were limited to 15-20%, and (2) results were more reasonable (e.g. Bureau of Public Roads - BPR function fit better) compared to taking the minimum or the average of F/C ratios. An underlying reason is the fact that the analysed sections are quite long ones with a length from 1.2 to 3 kilometres and they are strongly interdependent as it was observed that a heavily congested section can significantly influence the travel time on the whole route. It should be noted that the location of the maximum sectional F/C ratio is dynamically changing, which is quite natural.

Based on the distinguished traffic states it was eventually found that the fundamental diagram can also be used in an urban environment. However, there are some uncertainty concerning the transition states around the boundary of normal and congested states. After distinguishing measured F/C values based on whether those representing 'normal' or

'congested' traffic states, it was needed to transform congested ones. A conversion method was needed as observed F/C values that cannot be higher than 1 can describe congestion (e.g. a 0.7 F/C value can mean a slightly congested state), but in transport modelling congested states are measured with values above 1. Due to the validity of the speed-flow relationship, the transformation has been done using a "mirror" function which consists a contraction as well. The empirical background of the function comes from the difference between observed and modelled traffic states illustrated by Figure 3. A function that describes the conversion between observed and modelled saturation for the congested states was defined based on the VDF estimations of Juhász et al. (2016). It is presented by Equation (1):

$$F / C_{\text{mod}} = 1 + \frac{1 - F / C_{obs}}{c}, \tag{1}$$

where the modelled and observed saturation levels are represented by 'F/C_{mod}' and 'F/C_{obs}', and there is a correction (or contraction) parameter labelled by 'c'. Its value was calibrated around 1.2 during the VDF experiments.

4. Results and discussion

The process of the aforementioned saturation level correction resulted in a dataset consisting of a mean travel time and a standard deviation value for each F/C group. Based on the data a BPR function (a standard type of VDF, see Equation 2) was calibrated based on both 'normal' (not congested) and 'congested' states:

$$t = t_0 \cdot \left(1 + a \cdot \left(\frac{F}{C}\right)^b\right). \tag{2}$$

Within function (2) 't' is the mean travel time, while 't0' is the free-flow travel time. The estimated constant parameters are the following: a = 0.841, b = 2.52. During all estimations the dataset was grouped based on the F/C ratios in 5% intervals. Figure 4 illustrates the accuracy of the VDF estimation.

Based on the standard deviation values a function can be developed, which can describe the relationship between the standard deviation of travel times and F/C groups. Setting out from the shape, a standard cubic function turned out to fit the data points as three stages of the function can be observed. For very low F/C ratios the standard deviations of travel times are higher and decrease up to around the saturation level of 0.35 where the function has a local minimum. For higher F/C values the standard deviation is constantly increasing to the local maximum point (around 1.25 F/C value) from which there is a slight decrease. The reason is quite logical and well-described in the literature (see de Jong and Bliemer (2015) or Eliasson (2006)).

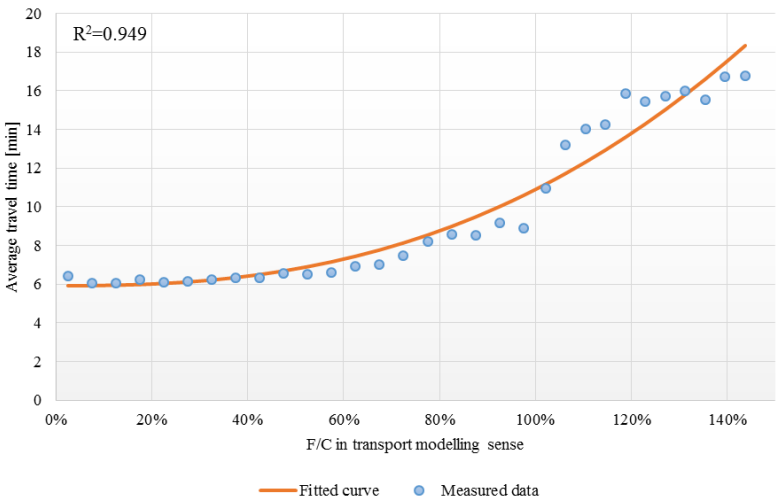


Fig. 4. Measured and modelled mean travel times (route no. 4)

For very low traffic volumes the traffic state could be 'instable' and the variability of travel times might be higher than normally expected. As traffic volume increases, the traffic state is becoming 'stable' up to the local minimum point. From this point the traffic starts to become heterogeneous and travel time variability is increasing with the saturation level. Then towards a heavy congestion state traffic is about to become homogeneous due to queuing, in which state the variability of travel times is decreasing. The process is also illustrated by the density plots of the travel time observations for specific saturation groups (see Figure 5).

One can note that the higher standard deviation values in case of very low saturation levels should be disregarded in the forecasting model as the higher variability of travel times is presumably coming from the higher level of freedom in choosing cruising speed. It means that this higher variability is reflecting a nearly free-flow traffic state in which the heterogeneity of car drivers is more perceptible. Despite the phenomenon is not correlated with congestion it is a feature of the proposed model. However, this issue can yield further considerations, analyses and possible modifications of the model.

Then standard deviation is described by Equation (3):

$$std = a \cdot \left(\frac{F}{C}\right)^3 + b \cdot \left(\frac{F}{C}\right)^2 + c \cdot \left(\frac{F}{C}\right) + d \cdot t_0 \,. \qquad (3)$$

As a result of a multiple linear regression analysis the estimated parameters are the following: a = -2.532, b = 6.515, c = -3.588 and d = 0.298. Figure 6 illustrates the accuracy the forecasting model of the standard deviation (TTR). It should be stressed that all measured data (travel times and standard deviation) were calculated for F/C groups (steps of 5%) as the forecasting problem was approached from a transport modelling point of view in which the saturation level (F/C ratio) has the largest influence. The standard deviation function has a point of inflection at around 0.85 F/C ratio, which seems to be theoretically appropriate as the boundary between ordinary and hyper-congestion should be somewhere around 1 but a lower value is also possible. It should be also stressed out, that the shape of the function comes from the above mentioned theoretical considerations (i.e. to adequately describe the phenomenon) and not because it provides the best fit to the data points.

Kouwenhoven and Warffemius (2016) suggests to not only calculate the raw standard deviation but to use a correction for the expected travel times. Then the deviation of the real travel times is calculated from the predicted travel times. In this study mean travel times and standard deviation values are calculated for F/C groups in which there can be data from different days and time periods which makes it impossible and unnecessary to determine expected travel time values. Therefore, this correction was not relevant for this research.

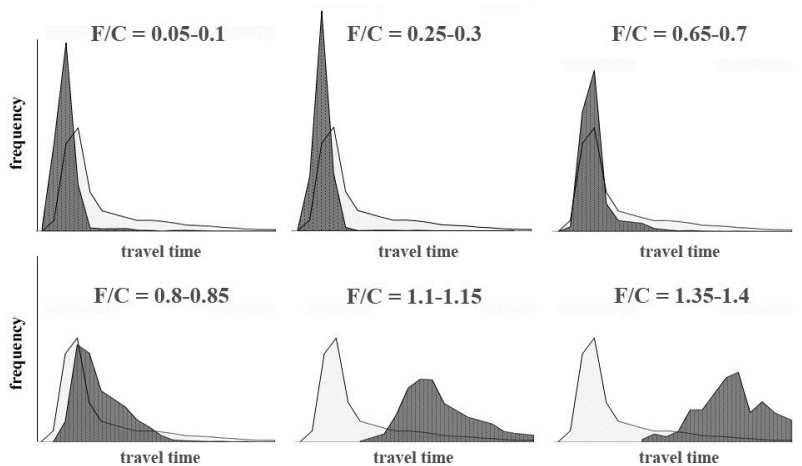


Fig. 5. Frequency of observed travel times for different F/C groups (route no. 4, dark grey – actual F/C group, light grey - total)

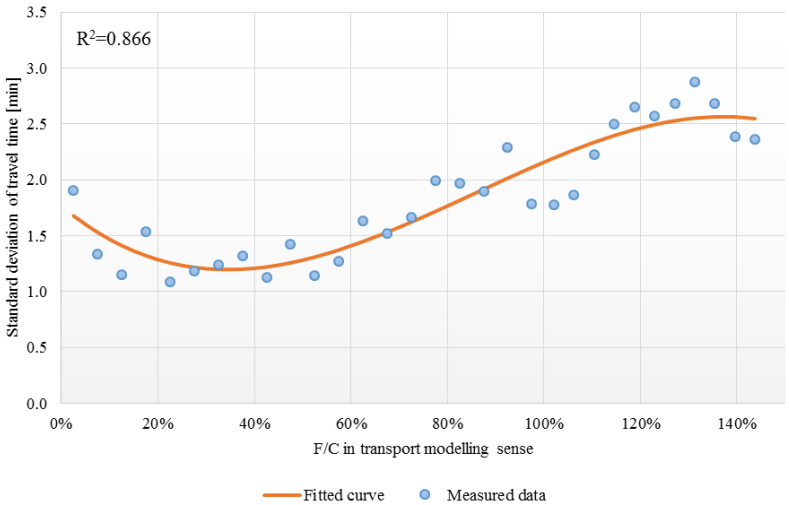


Fig. 6. Measured and modelled standard deviation of travel times (route no. 4)

The role of route length as an explanatory variable for standard deviation was also analysed but it was found that based on the data from Budapest it is not significant. However, other studies proved that it can also have an important role as on the one hand congestion is more likely and on the other hand delays could be compensated along a longer route. In this study there was not a big difference between the distances of the analysed routes and that might be a reason why length has not proved to be a significant factor.

Alternatively, other TTR forecast models were tested on the available dataset. This required another approach of analysis as most of the other methods are using mean delay (ratio of mean travel time and free-flow travel time) as an explanatory variable for standard deviation. The study of Eliasson (2006) seemed to be especially interesting to compare the results with. So based on its method, our daily data was split into 30-minute time periods. One can note that Eliasson used a 15-minute interval based on the implicit assumption that travellers base their decisions on this 'time resolution'. However, due to the measurement intervals previously mentioned, only a 30-minute split was possible. Applying this splitting, 672 data points could be created for the same 14 workdays we analysed before.

Analysing the relationship of absolute standard deviation and mean travel time as well as relative standard deviation (standard deviation divided by mean travel time) and relative increase in travel time (travel time divided by free flow travel time minus 1), the same findings can be found as by Eliasson (see Figure 7):

- standard deviation in absolute terms tends to increase with travel times;
- relative standard deviation increases with congestion but decreases for higher congestion levels.

A comparison of forecast methods was also carried out based on mean delays calculated by the calibrated VDF for each 5% F/C group as it would be normally measured during a project assessment. After some calibration model fit was adequate for all methods with R^2 values around 0.7. Only the NZTA model showed a lower value of 0.55. Results are illustrated by Figure 8 and they show that the method developed on the Budapest case has the best fit with a 1.5 sum of squared differences. The other models are slightly underestimating the standard deviation for lower F/C values and sums of squared differences are in the range from 1.9 to 2.2. It does not mean that the proposed model is universally better as the proposed model was designed based on the Budapest case and it is not surprising that it has the best fit. However, the results show that TTR can be forecasted based purely on the saturation level with similar accuracy to the existing models that predicts TTR based on mean-delay.

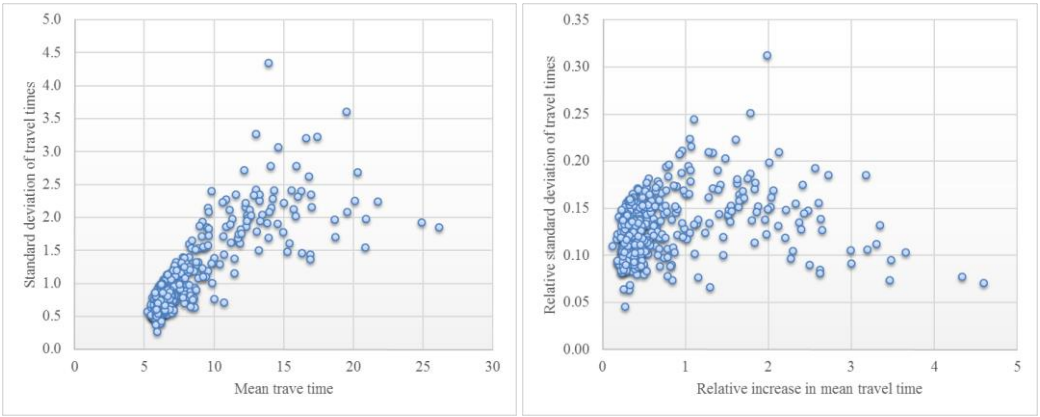


Fig. 7. The relationship of standard deviation and travel time (route no. 4) – each dot representing a 30-minute time interval

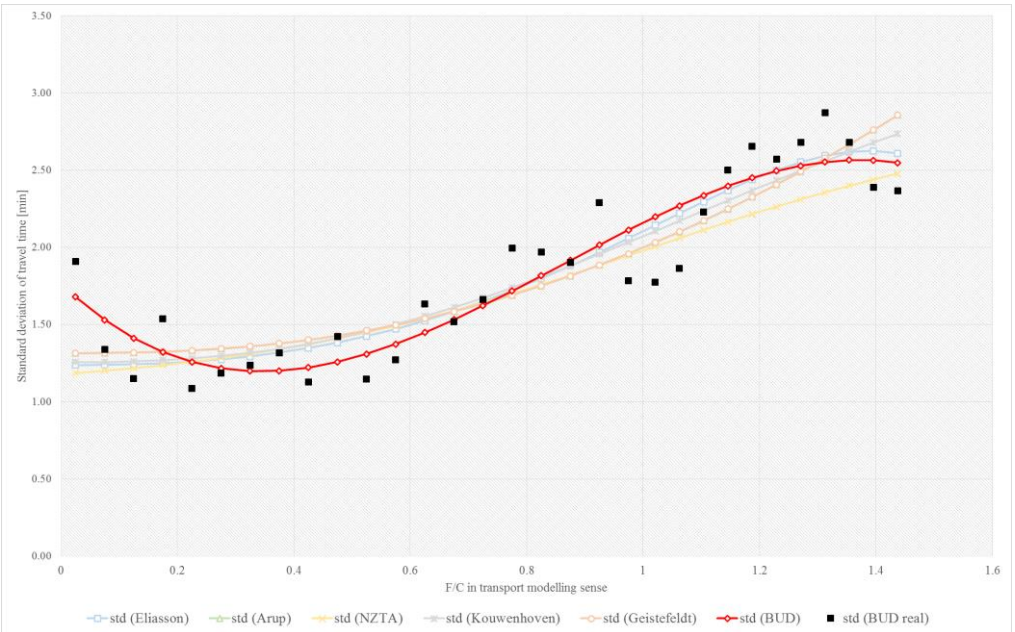


Fig. 8. Comparison of TTR forecast models (route no. 4)

5. Conclusion

Assessment of TTR as a fundamental factor in travel behaviour has become an important aspect in both transport modelling and economic appraisal. Improved reliability could provide a quite significant economic benefit if it is calculated in CBAs for which the theoretical background has already been set (definition of VOR). However, methods to forecast TTR as well as travel behaviour models including TTR effects are rather scarce and there is a need for development. Another important aspect could be the influencing factor of reliability in travel demand management and related policy-making as restrictive road projects (e.g. traffic calming projects) might decrease TTR in the whole transport system which besides the positive effects

of these interventions could mean an undesirable loss for the society. In addition to these, forecasting TTR might present new opportunities in the provision of real-time traffic information. Therefore, this paper aimed to analyse reliability, focusing exclusively on urban road transport as it is presumed that the issue is mostly significant in this setting. However, it is quite likely that assessment of TTR can be relevant in other settings such as long-distance or public transport trips as well.

This research pointed out that besides existing mean-delay-based models, TTR can be forecasted based on the volume-capacity ratio with adequate accuracy. The novelty of this result is that the issue of interdependence (endogeneity) of previous models can be resolved. Then it becomes possible to forecast TTR independently of travel time (or mean delay) which makes it easier to include TTR in travel behavioural models.

However, due to the limitations of travel time measurements and data (detailed in section 3.2) the proposed model and all of the results are based on route-level analyses with certain constraints (e.g. to use the maximum F/C value to characterise the saturation level of the route). In spite of the facts that (1) in another paper of the authors the validity of the speed-flow function was proved for urban routes and (2) other functions that forecasting TTR on a link-level provides very similar results, the proposed model is only valid for the route-level and not necessarily valid for its shorter sections (links) with different technical parameters. It should be noted that routes may consists of different link types and TTR at given F/C ratios might vary greatly across these types. However, the result of this paper suggests that the base model described in section 4 could still be used with proper calibration in appraising urban road projects.

Anyway, as a consequence of the shortcomings of this research, it should be also stressed that further analyses would need to check the validity and the universality of the results. Providing that sufficient data was available, it would be preferred to do the estimations on a link-level and to assess the difference between cities, road types, seasons, days, etc. Plus, a following further research might be able to develop TTR forecasting methods for urban public transport trips and cycling.

Acknowledgements

We are immensely grateful to the colleagues from Budapest Közút Plc. especially to Gergely Rónai for the help regarding the acquisition of data.

This research was supported by the EFOP-3.6.1-16-2016-00017 project.

References

[1] ARUP, 2003. Frameworks for Modelling the Variability of Journey Times on the Highway Network. London: Arup.

[2] BATES, J., POLAK, J., JONES, P., & COOK, A., 2001. The valuation of reliability for personal travel. Transportation Research Part E: Logistics and Transportation Review, 37 (2-3), 191-229.

[3] DAGANZO, C. F., 2007. Fundamentals of Transportation and Traffic Operations. Bingley: Emerald Group Publishing Ltd.

[4] DALE, H. M., PORTER, S., & WRIGHT, I., 1996. Are there quantifiable benefits from reducing the variability of travel times? Paper presented at the European Transport Conference 1996, Uxbridge, United Kingdom. Available from Internet: http://abstracts.aetransport.org/paper/download/id/464

[5] DE JONG, G. C. & BLIEMER, M. C. J., 2015. On including travel time reliability of road traffic in appraisal. Transportation Research Part A, 73, 80-95.

[6] ELIASSON, J., 2006. Forecasting travel time variability. Paper presented at the European Transport Conference 2006, Strasbourg, France. Available from Internet: http://abstracts.aetransport.org/paper/download/id/2491

[7] ESZTERGÁR-KISS, D. & RÓZSA, Z., 2015. Simulation results for a daily activity chain optimization method. Proceedings of the 4th International Conference on Models and Technologies for Intelligent Transportation Systems: 259-264. doi: 10.1109/MTITS.2015.7223265

[8] FICZERE, P., ULTMANN, Z., & TÖRÖK, Á., 2014. Time–space analysis of transport system using different mapping methods. Transport, 29 (3), 278-284.

[9] FOSGERAU, M. 2016. The Valuation of Travel Time Variability. Discussion Paper 04/2016, International Transport Forum.

[10] FOSGERAU, M., & HJORTH, K., 2008. The value of travel time variability for a scheduled service. Paper presented at the European Transport Conference 2008, Noordwijkerhout, the Netherlands.

[11] FOSGERAU, M., HJORTH, K., BREMS, C., & FUKUDA, D., 2008. Travel time variability - Definition and valuation. DTU Transport.

[12] GEISTEFELDT, J., HOHMANN, S., & WU, N., 2014. Ermittlung des Zusammenhangs von Infrastruktur und Zuverlässigkeit des Verkehrsablaufs für den Verkehrsträger Straße – Schlussbericht für Bundesministerium für Verkehr und digitale Infrastruktur. Ruhr Universität Bochum.

[13] GREENSHIELDS, B., 1935. A study of traffic capacity. Proceedings of the highway research board, 14 (1), 448-477.

[14] HORBACHOV, P., NAUMOV, V., KOLII, O., 2015. Estimation of the bus delay at the stopping point on the base of traffic parameters. Archives of Transport, 35(3), 15-26.

[15] ITF, 2010. Improving Reliability on Surface Transport Networks. [online] OECD Publishing. Available at: http://www.oecd-ilibrary.org/transport/improving-reliability-on-surface-transport-networks_9789282102428-en

[16] JUHÁSZ, M., KOREN, Cs., & MÁTRAI, T., 2016. Analysing the speed-flow relationship in urban road traffic. Acta Technica Jaurinensis, 9 (2), 128-139.

[17] JUHÁSZ, M., 2014. Assessing the requirements of urban traffic calming within the framework of sustainable urban mobility planning. Pollack Periodica, 9 (3), 3–14.

[18] KOUWENHOVEN, M., & WARFFEMIUS, P., 2016. Forecasting Travel Time Reliability in Road Transport. Discussion Paper 02/2016, International Transport Forum.

[19] KOUWENHOVEN, M., SCHOEMAKERS, A., GROL, R. V., & KROES, E. P., 2005. Development of a tool to assess the reliability of Dutch road networks. Paper presented at the European Transport Conference 2005, Strasbourg, France.

[20] LAIRD, J., NASH, C., & MACKIE, P., 2014. Transformational transport infrastructure: cost-benefit analysis challenges. Town Planning Review, 85(6), 709-730.

[21] MÁTRAI, T., 2013. Cost benefit analysis and ex-post evaluation for railway upgrade projects. Periodica Polytechnica Transportation Engineering, 41(1), 33–38.

[22] MÁTRAI, T., 2012. Cost benefit analysis and ex-post evaluation for railway upgrade projects – Ex-post economic evaluation, evaluation of traffic disturbance during construction and evaluation of travel time variability. MSc diss., Instituto Superior Téchnico.

[23] MÁTRAI, T., & JUHÁSZ, M., 2012. New Approach for Evaluate Travel Time Variability and Application for Real Case in Hungary. Paper presented at the European Transport Conference 2012, Glasgow, Scotland. Available from Internet: http://abstracts.aetransport.org/paper/download/id/3952

[24] NZ TRANSPORT AGENCY, 2010. Economic Evaluation Manual (Volume 1). Wellington: NZTA.

[25] ORTÚZAR, J. de D., & WILLUMSEN, L.G., 2011. Modelling transport. Chichester: John Wiley & Sons Ltd.

[26] PEER, S., KOOPMANS, C., & VERHOEF, E. T., 2010. Predicting Travel Time Variability for Cost-Benefit Analysis. Discussion Paper, Tinbergen Institute. 24 p. Available from Internet: http://papers.tinbergen.nl/10071.pdf

[27] RAO, A.M., & RAO, K. R., 2012. Measuring Urban Traffic Congestion – a Review. International Journal for Traffic and Transport Engineering, 2(4), 286–305.

[28] SPŁAWIŃSKA, M., 2015. Development of models for determining the traffic volume for the analysis of roads efficiency. Archives of Transport, 35(3), 81-92.

[29] STAMOS, I., MARIA, J., GRAU, S., MITSAKIS, E., & MAMARIKAS, S., 2015. Macroscopic fundamental diagrams: simulation findings for Thessaloniki's road network. International Journal for Traffic and Transport Engineering, 5, 225–237.

[30] SUSILAWATI, S., TAYLOR, M. A. P., & SOMENAHALLI, S. V. C., 2013.

Distributions of travel time variability on urban roads. Journal of Advanced Transportation, 47(8), 720–736.

[31] TAYLOR, M. A. P., 2013. Travel through time: the story of research on travel time reliability. Transportmetrica B, Transport Dynamics 1(3), 174-194.

[32] TRB, 2011. Travel Time Reliability. TRB Transportation Economics Committee. Available at: http://bca.transportationeconomics.org/benefits/travel-time-reliability [Accessed 01 November 2011]

[33] UNITED NATIONS, 2013. World Population Prospects The 2012 Revision Volume I: Comprehensive Tables. New York: UN.

[34] UNITED NATIONS, 2015. World Urbanization Prospects: The 2014 Revision. New York: UN.

[35] VASVÁRI, G., 2015. Additive Effects of Road Functions. Periodica Polytechnica Civil Engineering, 59(4), 487–493.

[36] VÖRÖS, T., JUHÁSZ, M., & KOPPÁNY, K., 2016. The measurement of indirect effects in project appraisal. Transportation Research Procedia, 13, 114-123.

[37] WOOLLETT, N., VAUGHAN, B., & LUNT, G., 2015. Re-validation of speed/flow curves. Paper presented at the European Transport Conference 2015, Frankfurt, Germany.

ARCHIVES OF TRANSPORT www.czasopisma.pan.pl PAN POLSKA AKADEMIA NAUK www.journals.pan.pl
Volume 43, Issue 3, 2017
ISSN (print): *0866-9546*
e-ISSN (online): *2300-8830*
DOI: *10.5604/01.3001.0010.1795*

A NEW SIMULATION-OPTIMIZATION APPROACH FOR THE CIRCULATION FACILITIES DESIGN AT URBAN RAIL TRANSIT STATION

Afaq Khattak[1], Yangsheng Jiang[2], Juanxiu Zhu[3], Lu Hu[4]

[1,2,3,4] Traffic Engineering Department, School of Transportation and Logistics, Southwest Jiaotong University, National United Engineering Laboratory of Integrated and Intelligent Transportation, Chengdu, Sichuan China

[1]e-mail: af.transpo@gmail.com
[2]e-mail: jiangyangsheng@swjtu.cn
[3]e-mail: zhjuanxiu@163.com
[4]e-mail: hulu361@126.com

Abstract: *Width design of the urban rail transit stations circulation facilities is a vital issue. The existing width design approach failed in fully considering the essential factors such as fluctuation in passengers' arrival process, fluctuation and state-dependence in passengers walking speed and the blocking when passengers' demand exceeds the capacity of facilities. For this purpose, a PH-based simulation-optimization approach is proposed that fully considers the fluctuation, the state-dependence, Level of Service (LOS) and blocking effect. This novel approach provides automatic reconfiguration of the widths of circulation facilities by a concurrent implementation of a PH-based Discrete-Event Simulation (DES) model and the Genetic Algorithm (GA). The proposed PH-based simulation- optimization approach and the existing design approaches based on the exponential and deterministic models are applied to design the widths of circulation facilities. The results reveal that the circulation facilities designed by the proposed approach have larger widths. Similarly, increase in the SCV of arrival interval results in increasing the widths designed by the proposed approach increase while the widths of the other two approaches stay the same. The width designed of the proposed approach increase at faster rate than that of the other two approach when the passengers' arrival rate increases.*

Key words: *Urban Rail Transit Station, Circulation Facilities, PH-based Discrete-Event Simulation, Genetic Algorithm, PH-based Simulation-Optimization.*

1. Introduction

The urban rail transits are playing a significant role in the urban transport, especially in metropolises. The urban rail transit stations are the operational systems consisting of a framework of infrastructures, service facilities, and personnel; they are the points of connection between arrivals and departures of passengers. In recent years, investment and improvement in the urban rail transits encouraged the people to switch from driving to transits.

The performance of the urban rail transit station service facilities naturally became a great concern to both passengers and operators. The better performance of these service facilities is the reflection of enhanced design while the inadequate design often leads to high-level congestion, the longer travel time of passengers between the service facilities, inefficient space utilization, resource wastage and increase in the waiting time of passengers which in turn implies that there is a direct correlation between design and performance. The width (W) of the circulation facilities (corridors and stairs) is a most significant factor and its design is a vital issue. It is obtained by using the passengers' arrival rate divided by the service rate (flow rate) per unit width under a given Level of Service (LOS) in the Transit Capacity and Quality of Service Manual (TCQSM) (Kittelson et al., 2003) but they have a several shortcomings, such as;

- The design procedure neglect fluctuation in passengers' arrival process.
- The fluctuation as well state-dependence walking speed of passengers is ignored.

- Several service facilities of urban rail transit stations are designed separately and the correlation between them is fully neglected.
- The analysis and design procedure neglect blocking phenomenon in different facilities when the passengers' demand exceeds the serviceability of the facilities.

Due to all these shortcomings, the circulation facilities designed by the TCQSM always show poor performance and face blockage even during the off-peak hours. The heavy congestion and blocking can cause serious accidents if not controlled. Thus, there is an urgent need for a new design approach for circulation facilities that overcome the shortcomings.

Therefore, the study reported in this research details the Discrete-Event Simulation (DES) as well as the *simulation-optimization* approach for the analysis and optimal design of urban rail transit station service facilities, considering both the fluctuation in passengers' arrival process and the service times of the circulation facilities. The Phase-Type (PH) distribution considers the randomness factor and therefore it is used to fit the passengers' arrival and service processes in the DES model. Moreover, the PH-based *simulation-optimization* approach, integrating the PH-based DES models of the service facilities and the optimization algorithm based on Genetic Algorithm (GA) is used to design the facilities and eliminates the need to solve explicit analytical expressions over a large time span, as in the case of mathematical optimization.

The assessment of LOS in circulation facilities uses the area occupied per passenger (m^2/ped) as the basis for classification (See Exhibit 7-3 and 7-7 in Reference Kittelson et al. 2003). It reflects proximity to other passengers and is therefore considered as an indicator of the passenger level of comfort and freedom to maneuver without conflict. In this paper, both the corridors and stairs are designed under the LOS 'B'. According to TCQSM, the minimum LOS 'B' values for corridors and stairs are 2.3 m^2/ped and 1.4 m^2/ped respectively.

Moreover, in this research, the Genetic Algorithm (GA) and PH-based BES are implemented in the *MATLAB*® Scientific Computing Environment and *SimEvents*® simulation software (a Discrete-Event Simulator in the *MATLAB*®/*Simulink*® family), respectively.

- The *MATLAB*® offers a computational environment for optimizing hybrid discrete-event and time-based models, that allows for a great flexibility in scripting and modifying the optimization objective and constraint functions. It also making easier to tie together the parallel DES and optimization script without the pain of the context transferring into the multiple softwares.
- Just like other simulation tool, such as Arena , Extend, Witness) and Any Logic., the *SimEvents*® (Banks, 2010) allows the representation of complex Discrete-Event Systems by a network of queues, servers, gates and switches based on the events. Its integration with the *MATLAB*® simplifies the modeling process of the hybrid dynamical systems, which include discrete-time, continuous time and discrete-event systems. The *SimEvents*® contains libraries and block sets that model the basic components of DES. By inter-connecting these building blocks, one can easily model a DES of transportation systems, communication networks, and manufacturing systems, etc.

2. Literature Review

Several researches has been carried out to devise the new width design approach for the circulation facilities in urban rail transit stations as well as other buildings such as residential, hospitals and universities. Due to the inherent characteristics of circulation facilities, such as the relationship between the facilities and passengers (servers and customers), the fluctuation and state-dependence in the passengers' flow, many researches modeled the circulation facilities as various queuing systems. Based on this, both the analytical and simulation models are developed. The first approach uses mathematical techniques often called queuing analytical models to estimate the performance measures by using mathematical equation systems. The second approach is a computer simulation of the facilities. In the simulation environment, all quantities can be readily observed and the parameters can be changed to examine their influence on the system.. Generalized M/G/C/C state-dependent analytical queueing models pedestrian traffic flow established by Yuhaskiet al. (1989), Smith et al. (1991), Cheah et al. (1994), Cheah and Smith (1994) and Chen et al. (2012). Similarly, Jian and MacGregor Smith (1997)

developed a queuing model for the vehicular traffic flow. Vandaele et al. (2000) developed a finite capacity queuing networks to consider traffic flow studies on roads. Mitchell and MacGregor Smith (2001) extended their work to analyze and design the series, splitting and merging topologies of pedestrian network by using an analytical approximation methodology. Cruz et al. (2005) developed a state-dependent M/G/C/C queueing networks to determine the optimal capacity and number of servers. Jiang et al. (2010) modelled the urban rail transit station corridor facility as a M/G/1 queuing system wth the passenger arrival process based on exponential distribution and a service time based on general random distribution. Bedell and Smith (2012) examined the combination of multi-server and state-dependent M/G/C/K, M/G/C/C queues in transportation and material handling systems. Xu et al. 2014 analyzed the Urban rail transit station Capacity (SSC) as M/G/C/C state-dependent queuing network. A new concept according to the gathering and scattering process was defined.

With the advancement of computer technologies, the simulation approach has been emerged and many researchers focused on simulation approach for the analysis and design purposes. The G/M/1 queuing network simulation model by Lovas (1994), the M/G(*n*)/C/C state dependent network simulation model by Cruz et al. (2005) and Khalid et al. (2013). Ying et al. (2014) developed a queuing simulation and optimization model for number of ticket windows at urban rail transit station. A DES model is also developed by (Jiang and Lin, 2013) for the evaluation and optimization of the Ticket Vending Machines (TVM) at urban rail transit station using log-normal distribution and gamma distribution for arrival and service processes, respectively. In these researches, queuing systems are translated into Discrete-Event Simulation (DES) models. Based on the DES models, both evaluation and optimization are carried out. However, circulation facilities description still needs to be improved in the above researches.

Besides DES, another simulation category is also well known, that is the microscopic simulations. Microscopic simulation models are elaborate as they depict individual characteristics and behaviors of the pedestrians (Teknomo et al., 2006, Kaakai et al., 2007) as well as transportation system (Jacyna et al., 2014). However, they require extensive calibration work and larger computation time at the same time. On the contrary, DES does not require the specific physical environment and passenger entity, making it more efficient and easier to calibrate than microscopic simulation models. Therefore, DES is taken as an efficient and accurate simulation method with a wide range of application (Hassannayebi et al., 2014). Another advantage of the DES is that *simulation-optimization* can be carried out conveniently based on DES models due to its universality and efficiency.

From the review of advanced stochastic processes study, we found that the Phase-Type (PH) distribution has substituted the exponential distribution in several fields including; healthcare, queuing systems, manufacturing processes and communication systems. The reason to use PH distribution for fitting the arrival interval and service time in queuing system is its own apt analyticity, universality, and computability Jiang et al. (2013). Theoretically, it can be fit to any positive random number infinitely which has resulted in the emergence of ample PH-based queuing models including PH/PH/1 by Krishnamoorthy et al. (2008) and PH/PH/1/C by Alfa and Zhao (2000). In the transportation domain, Hu et al. 2013 for the first time applied the PH distribution to fit the passengers' flow arrival interval distribution at urban rail transit station which has revealed a good data fitting effect. It has opened the ways for using PH distribution in the field of traffic and transportation. Reijsbergen et al. (2015) proposed a methodology of constructing stochastic performance model for public transportation network using PH distribution.

Hu et al. (2015) presented an analytical PH/PH/C/C state-dependent queuing model for the analysis and design of urban rail transit station corridors. The PH/PH(*n*)/C/C state dependent queuing model take the state-dependence in service time into consideration. State dependence describes the phenomenon in circulation facilities the number of passengers (referred as system state *n*) affect the walking speed, which eventually affects the service time. However, it is very difficult to solve the PH/PH(*n*)/C/C state dependent analytical model even for a single facility. The complexity of solving the PH/PH(*n*)/C/C state dependent network model will be much larger due to the matrix operations. In

addition, the blocking probability is not controlled when designing the width for the single corridor facility in Hu et al. (2015).

Recently, *simulation- optimization* has become a popular and efficient tool in many domains (Banks, 2010; Hagendorf et al., 2013; Jiang et al., 2013 and Jiang et al., 2015). It involves the optimization of model inputs by using simulation for the computation of parameters (Figueira et al., 2014). Therefore, it is not necessary to provide an explicit analytical expression of the objective or constraint functions for optimization as in the case of analytical approach (Swisher et al., 2000; Fu, 2002, Cassandras et al., 2009; Hagendorf et al., 2013). This is especially useful in some practical situations where the explicit analytical formulas are too complex to be deduced.

To find the optimal widths for the circulation facilities by a PH-based *simulation-optimization* aapproach, we need to implement an optimization approach integrated with PH-based DES model. The Genetic Algorithm (GA) is chosen in this paper. There are several reasons for applying a GA rather than any other traditional optimization methods. One of the important reasons is its implicit parallelism (Swisher et al., 2000; Hubscher-Younger et al., 2012; Messac, 2015 and Lewczuk, 2015). The GA searches parallel from a population of points. As GA has multiple offspring, it can explore the solution in different directions at a time giving it greater chance to find the optimal solution, while other traditional methods search from a single point and may trap in local optimal solution.

Based on the above analysis, we aim to propose a new PH-based *simulation-optimization* approach for the width design of circulation facilities. The contribution of this paper falls into two aspects. First, we establish a PH-based DES model to describe the circulation facilities (include stairs and corridors) in the urban rail transit station. The PH-based DES model captures the general fluctuation in passengers' arrival and service facilities. It also takes the state-dependence in service time into consideration. Therefore, it can be used to accurately evaluate the performance of the circulation facilities. Besides, it also serves as an important tool to validate the PH/PH(*n*)/C/C analytical model developed in Hu et al. (2015). Second, we develop a PH-based *simulation-optimization* approach by implementing the PH-based DES model and the GA to work concurrently. The PH-based *simulation-optimization* approach determines the optimal widths of circulation facilities by considering the requirements on both LOS and blocking probability. Therefore, the circulation facilities designed by it enjoy higher service quality and less congestion. The proposed PH-based *simulation-optimization* approach can support decision making in circulation facilities design

3. Notations

Notation	Description
α	Initial probability vector
D	Transient Generator Matrix
L	Length of the walkway (m)
W	Effective width of the walkway (m)
C	Capacity of the walkway
n	Number of passengers (system state)
λ	Passenger arrival rate (ped/h)
c_a^2	Squared Coefficient of variation of arrival rate
q	Peak-hour volume
ε	Peak-hour factor
h	Mean headway between the trains (sec)
c_h^2	Squared coefficient of variation of headway
μ_n	State-dependent service rate
T_n	State-dependent service time of walkway
V_n	State-dependent walkway speed of passengers
$c_{s,n}^2$	State-dependent squared coefficient of variation of walkway service rate
P_c	Blocking probability
ES	Mean area occupied per passenger
f	Degree of Erlang distribution
U	Uniformly distributed random number

4. Definition of PH Distribution

Before going into the details of PH-Based DES model, the PH distribution is discussed first. The PH distribution is a probability distribution that represents the time to absorption in a Continuous-Time Markov Chain (CTMC) with one absorbing

state and all the other transient states (Neuts, 1981). PH distributions are commonly represented by the pair (α , **D**). Here, α is an initial probability vector and **D** is a transient generator matrix as follows

$$\alpha = (\alpha_1, \ldots\ldots \alpha_n), \quad \mathrm{D} = \begin{pmatrix} d_{11} & \cdots & d_{1n} \\ \vdots & \ddots & \vdots \\ d_{n1} & \cdots & d_{nn} \end{pmatrix}.$$

The probability density function (PDF) and cumulative distribution function (CDF) of PH distribution are given by Equation (1) and (2)

$$f(x) = \alpha e^{\mathrm{D}x} d \tag{1}$$

$$F(x) = 1 - \alpha e^{\mathrm{D}x} 1 \tag{2}$$

where: d = **D.1** and **1** is a column vector of one's of the appropriate size .

There are four conditions given for fitting PH distribution (Sadre and Haverkort, 2011; Sadre, 2007) based on the mean value and the SCV:

1) If the SCV c^2 for both the arrival and service processes is less than 1, a hypo-exponential distribution is used to fit the arrival and service processes with the number of phases given by $m = \frac{1}{c^2}$, the initial probability vector is $\alpha = (1, 0, \ldots., 0)$ and the matrix **D** is expressed by:

$$\mathbf{D} = \begin{pmatrix} -d_0 & d_0 & & & & \\ & -d_1 & d_1 & & & \\ & & .. & .. & & \\ & & & & -d_{m-2} & d_{m-2} \\ & & & & & -d_{m-1} \end{pmatrix},$$

where:

$$d_j = \frac{m}{E[X]} \text{ for } 0 \le j < m-2;$$

$$d_{m-1} = \frac{2m\left[1 + \sqrt{\frac{1}{2}m(mc^2 - 1)}\right]}{E[X](m + 2 - m^2c^2)};$$

$$d_{m-2} = \frac{m\lambda_{m-1}}{2\lambda_{m-1}E[X] - m}.$$

2) If the SCV c^2 is greater than 1 for both the arrival and service process, a hyper-exponential distribution is used for fitting with the number of phases $m = 2$, the initial probability vector is $\alpha = (g, 1-g)$ and the matrix **D** is given by :

$$\mathrm{D} = \begin{pmatrix} \frac{-2g}{E[X]} & 0 \\ 0 & \frac{-2(1-g)}{E[X]} \end{pmatrix} \text{ and } g = \frac{1}{2} + \frac{1}{2}\sqrt{\frac{c^2 - 1}{c^2 + 1}}$$

3) If c^2 is equal to 1, then the approximation corresponds to an Exponential distribution.
4) If c^2 is very small i.e., $c^2 \le 1/30$ then the PH distribution with a large number of states is obtained and its approximation corresponds to an Erlang-30 distribution.

Jiang et al. (2013) and Hu et al. (2015) have achieved a good fitting effect for the passenger arrival interval from the train as well state-dependent service time of circulation facilities by using a PH distribution with any SCV. The four conditions show that we can determine the PH representation for the arrival interval and service time based on λ_i, $c_{i,a}^2$, $\mu_{i,n}$ and $c_{i,s,n}^2$. Note that $E[X]$ is the reverse of the arrival rate λ_i and the service rate $\mu_{i,n}$.

5. Circulation Facilities as a Queuing System

The necessary assumptions used in this paper are discussed first followed by describing the PH-based DES model of circulation facilities.

5.1. Assumptions

Few basic assumptions are presented before the modeling of circulation facilities.

- The circulation facilities including both corridors are stairs are rectangular in shape with Length (L) and Width (W). The Width W is the effective width of circulation facility and the total width is obtained by adding a buffer of 0.5m on each side to the effective width.
- The passengers are assumed to be uniformly distributed in the circulation facilities. This is quite rare from a practical point of view but an important assumption for queuing analysis which

is used in many relevant studies such as Yuhaski et al. (1989), Jiang et al. (2015) and Hu et al. (2015).

- Only the alighting passenger flow from the train is considered. The proposed approach can also deal with bi-directional or multi-directional passenger flow by changing some parameters as well using additional blocks of *SimEvents*® simulation software.

5.2. Modeling of Circulation Facilities

The circulation facility of urban rail transit station is a type of open queuing network. Passengers enter the stairs or corridors and leave the facilities after receiving services. The circulation facilities include stairs and corridors (see Figure 1a) and they are turned into a topology of the queuing network system (see Figure 1b). The circulation facilities (nodes of a queuing network) are designated by $i = 1,2.....N$, where; N is the total number of circulation facilities.

The flow lines represent the passengers flow at different circulation facilities with the routing probabilities represented by R_{st}. Here 's' is the preceding facility and 't' is the successor facility. When the alighting passengers on the platform entering into a circulation facility, they occupy the spaces in the facility (squares) (See Figure 2). Each available space in the circulation facility acts as a server (service desk). The passengers spend some time (walking/travel time) in the circulation facility and then exit. The passengers and the circulation facility can be viewed as a queuing system with passengers as customers, the spaces in the circulation facility as servers and the process of walking in the circulation facility as a service process.

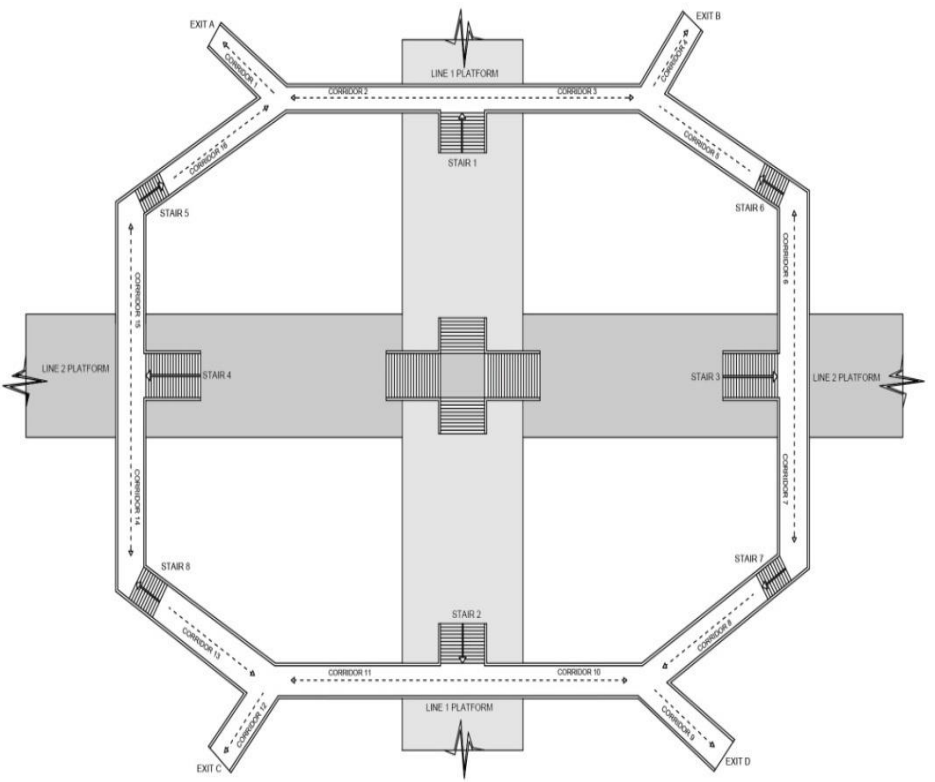


Fig. 1a. Queuing network representation of circulation facilities - Layout of the urban rail transit station circulation facilities

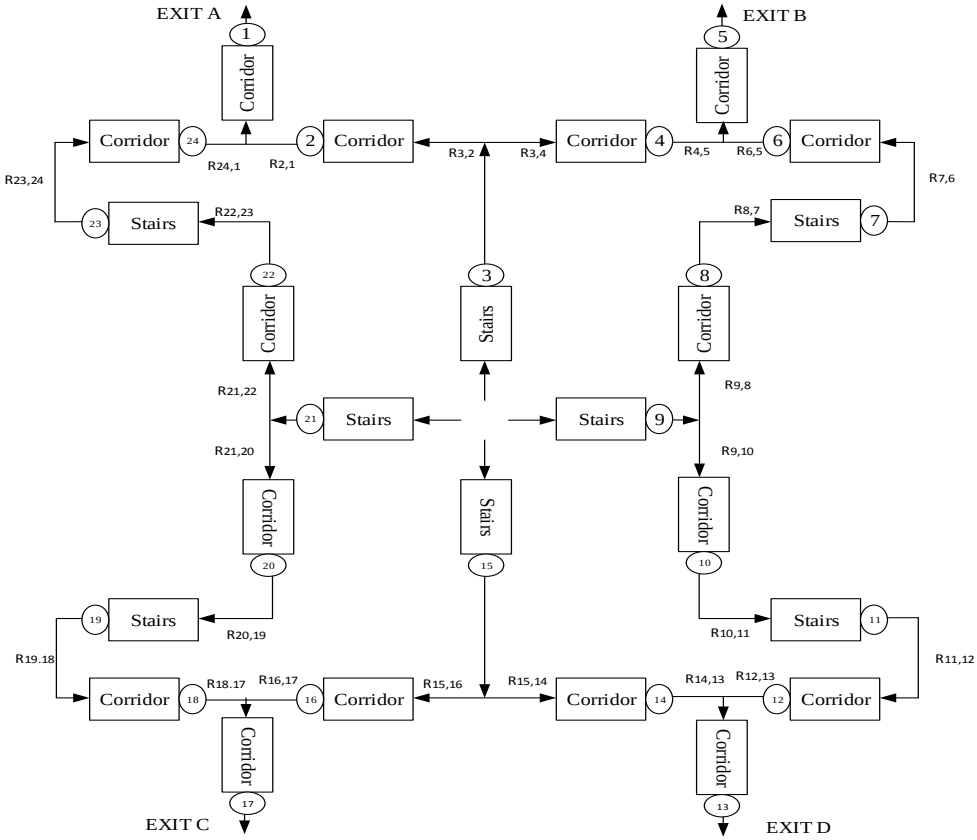


Fig. 1b. Queuing network representation of circulation facilities - Queuing network topology for circulation facilities

The number of passengers 'n' changes in the circulation facility dynamically over time. As the number of passengers in the circulation facility increases, the slower passengers block faster passengers. Thus, higher passenger densities reduce the individual passenger's walking speed. The speed is reduced to 0 when the number of passengers n reaches the capacity of the circulation facility $C = 5LW$, which means the passenger flow in the circulation facility can be viewed as stopped when the density of passengers is a 5 ped/m^2 (Tregenza, 1976). The phenomenon of variation in walking speed with the increase or decrease in the number of passengers 'n' in the circulation facility is known as state-dependence. Hence, any circulation facility can be described as a state-dependent queuing system with passenger arrival interval represented by the random variable A_i, state-dependent service time of the circulation facility $B_i(n)$, the number of servers (available positions) C_i, i.e., a $A_i/B_i(n)/C_i/C_i$ queuing system.

Since the value of C_i is generally very high in hundreds and even thousands. The queuing systems with a high value of C_i are difficult to simulate and cause serious problems in optimization such as low optimization efficiency. Therefore, it is necessary to simplify the $A_i/B_i(n)/C_i/C_i$ queuing system. We use the idea of transformation which is also used in relevant researches (Jiang et al., 2015; Hu et al., 2015). The transformation works as follows. A virtual line and a virtual server are set at the exit of the circulation facility, as shown in Figure 2.

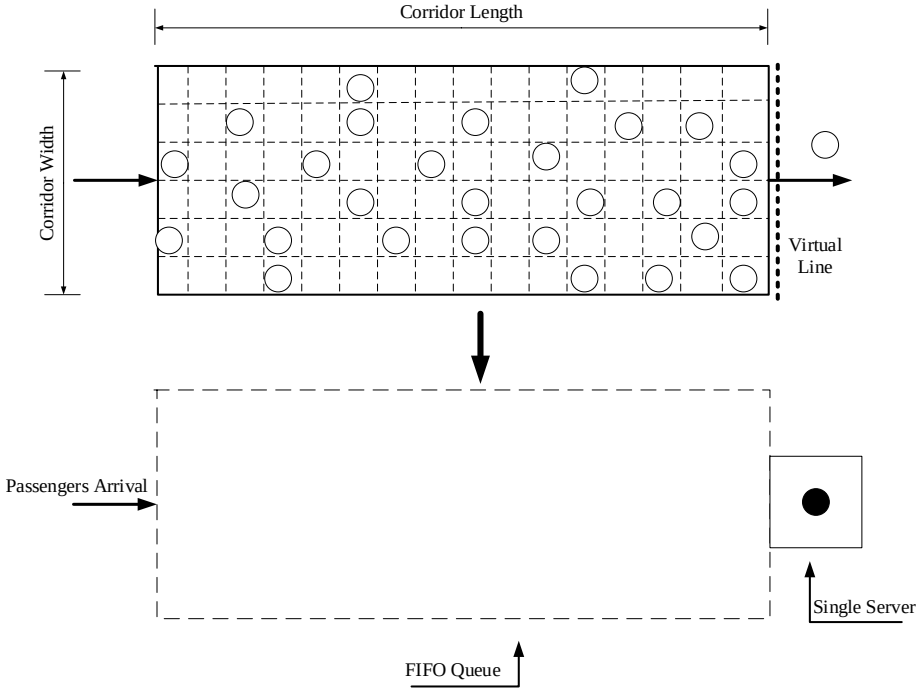


Fig. 2. Transformation of circulation facility to a single server queuing system

When two sequential passengers pass through the virtual line, the time interval T_n is recorded. If the time at which the previous passenger leaves the circulation facility is viewed as the time the next passenger begins to be served, then the time interval between these two sequential passengers passing through the virtual line is equal to the service time of the virtual server. In this way, the $A_i/B_i(n)/C_i/C_i$ queuing system with C_i parallel-serial servers can be transformed equivalently to an $A_i/B_i'(n)/1/C_i$ queuing system with a single server. Note, the service time of the virtual server $B_i'(n) = B_i(n)/n$. How to calibrate the parameters for the PH arrival interval and service time will be discussed in the next subsection.

5.3. Passengers' Arrival Process

The passengers' arrival process to the i^{th} circulation facility of urban rail transit station is specified by passenger the arrival rate λ_i and squared coefficient of variation (SCV) of arrival interval $c_{i,a}^2$ (Jiang et al., 2013). During the planning and design phase of urban rail transit station circulation facilities, the peak hour volume (q) and the peak-hour factor (ε) are usually given. So we can calculate λ_i and SCV ($c_{i,a}^2$) of the i^{th} circulation facility by:

$$\lambda_i = \frac{q}{3600\varepsilon} \tag{3}$$

$$c_{i,a}^2 = \frac{e^{6.819\varepsilon}(\varepsilon - 1)^2}{4\varepsilon - 1} \tag{4}$$

If the mean headway (h) between trains and the squared coefficient of variation of headway (c_h^2) is also given (for the existing urban rail transit station), then $c_{i,a}^2$ can also be calculated by:

$$c_{i,a}^2 = e^{0.503c_h^2}\left[\left(\frac{qh}{3600\varepsilon}\right) - 1\right] \tag{5}$$

5.4. State-dependent Service Phase

According to TCQSM (Kittelson et al., 2003) and the traffic flow theory, the passengers flow rate (μ) is given by Equation (6):

$$\mu = kV \tag{6}$$

Here k is the density of passengers and V is the passengers' walking speed in the circulation facility. In the case of urban rail transit station circulation facilities, the passenger flow rate is the number of passengers passing through the circulation facility per unit time. The reciprocal of flow rate $1/\mu$ is referred as the time interval of the passengers leaving the circulation facility which is also the state-dependent service time T_n of the single virtual server in Figure 2. Therefore, the state-dependent service time of the i^{th} circulation facility can also be expressed as:

$$T_{i,n} = 1/\mu_{i,n} = L_i / nV_{i,n},\ i = 1,2.....N \tag{7}$$

The state-dependent service rate of the i^{th} circulation facility can be written as:

$$\mu_{i,n} = 1/T_{i,n} = nV_{i,n} / L_i,\ i = 1,2.....N \tag{8}$$

Here L_i is the length and $V_{i,n}$ is the state-dependent walking speed of passengers passing through the i^{th} circulation facility. Yuhaski et al. (1989) developed an exponential model to describe the state-dependent walking speed in i^{th} circulation facility, shown by Equation (9):

$$V_{i,n} = V_1 \exp\left[-\left(\frac{n-1}{\omega_i}\right)^{\gamma_i}\right],\ i = 1,2.....N \tag{9}$$

where:

$$\gamma_i = \ln\left[\frac{\ln(v_{i,a}/v_{i,1})}{\ln(v_{i,a}/v_{i,1})}\right] \Big/ \ln\left(\frac{a_i - 1}{b_i - 1}\right),$$

$$\omega_i = (a_i - 1)\left[\ln\left(\frac{v_{i,1}}{v_{i,a}}\right)\right]^{1/\gamma_i}.$$

Thus, the Equation (9) can now be written as:

$$\mu_{i,n} = nV_1 \exp\left[-\left(\frac{n-1}{\omega_i}\right)^{\gamma_i}\right] / L_i,\ i = 1,2.....N \tag{10}$$

In order to consider the randomness of service time in the i^{th} circulation facility, the squared coefficient of variation (SCV) of service time should be taken into account. The state-dependent SCV of service time ($c^2_{i,s,n}$) for the i^{th} circulation facility is given by:

$$c^2_{i,s,n} = \left[\left(\frac{\delta_{i,1}}{v_{i,1}}\right)\exp\left(\left(\frac{n-1}{\omega_i}\right)^{\gamma_i} - \left(\frac{n-1}{\omega'_i}\right)^{\gamma'_i}\right)\right]^2 \tag{11}$$

where,

$$\gamma'_i = \ln\left[\frac{\ln(\delta_{i,a}/\delta_{i,1})}{\ln(\delta_{i,a}/\delta_{i,1})}\right] \Big/ \ln\left(\frac{a_i - 1}{b_i - 1}\right),$$

$$\omega'_i = (a_i - 1)\left[\ln\left(\frac{\delta_{i,1}}{\delta_{i,a}}\right)\right]^{1/\gamma_i}.$$

$v_{i,1}$ - Mean walking speed when there is only one passenger in the i^{th} circulation facility.

$\delta_{i,1}$ - Standard deviation of walking speed when there is only one passenger in the i^{th} circulation facility.

$v_{i,a}$ - Mean walking speed when there are $a_i = 2L_iW_i$ passengers in the i^{th} circulation facility.

$\delta_{i,a}$ - Standard deviation of walking speed when there are $a_i = 2L_iW_i$ passengers in the i^{th} circulation facility.

$v_{i,b}$ - Mean walking speed when there are $b_i = 4L_iW_i$ passengers in the i^{th} circulation facility.

$\delta_{i,b}$ - Standard deviation of walking speed when there are $b_i = 4L_iW_i$ passengers in the i^{th} circulation facility.

After fitting the PH distribution, the passenger arrival process can be described by the initial probability vector α_i and the transient generator matrix D_i as:

$$A_i \sim \mathrm{PH}(\alpha_i, D_i) \quad i = 1, 2.....N$$

The state-dependent service process of the i^{th} circulation facility can be described by the initial probability vector $\beta_{i,n}$ and the transient generator matrix $H_{i,n}$ as:

$$B_i(n) \sim \mathrm{PH}(\beta_{i,n}, H_{i,n}) \quad i = 1, 2...N \quad \text{and } n = 1, 2....C$$

The above initial probability vectors and transient generator matrix will be used for generating the PH random variates in the PH-based DES model.

6. PH-based DES Model of the Circulation Facilities

First, we introduce the generation of PH random variates that are the key ingidient for PH-based DES model. Then a PH-based DES model of circulation facilities is developed in the *SimEvents*® to evaluate the performance measures of the circulation facilities.

6.1. Generation of PH Random Variates

PH distribution is proposed in this paper to simulate the passengers' arrival rate and state-dependent service time of circulation facilities. Neuts (1981) developed a 'Count Procedure' for the efficient generation of PH random variates relies on generating an Erlang-distributed sample with degree f and parameter φ given as:

$$\mathrm{Erl}(f, \varphi) = -\frac{1}{\varphi} \ln\left(\prod_{j=1}^{f} U_j \right)$$

A pseudo-codes description of generating the PH random variates in this work is as follows:

Pseudo-code 1.

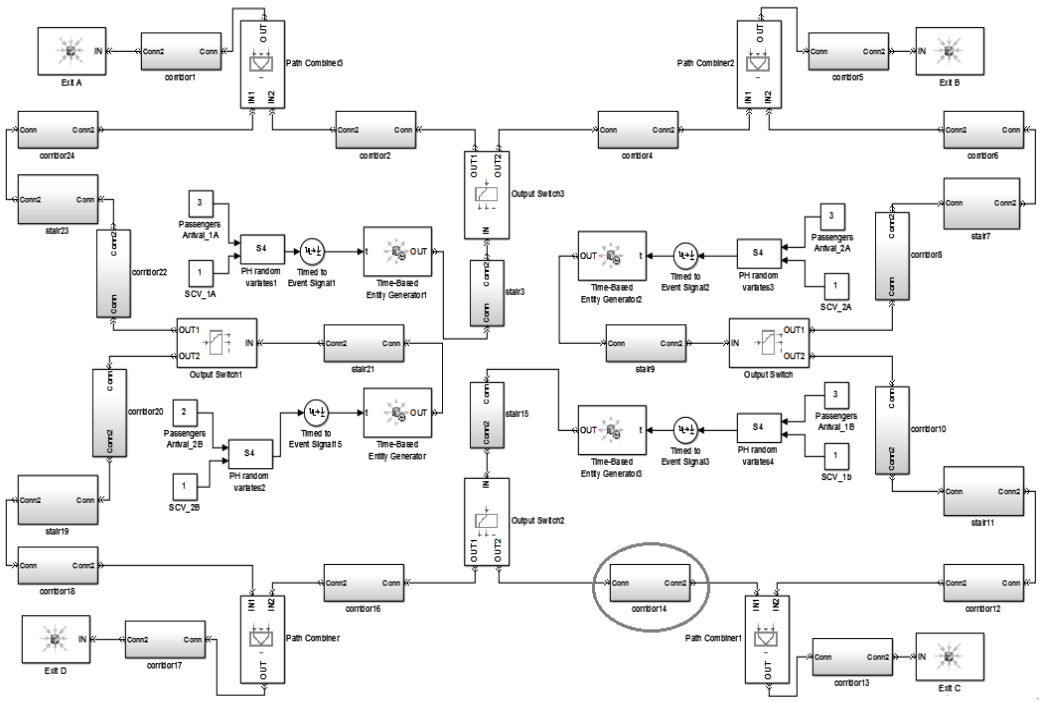
1) $x_{ph} := 0, f_j = 0, \text{for } j = 1, 2.....n$ Draw an α-distributed discrete sample for the initial state.
2) The chain in the state j,
   i. $f_j += 1$
   ii. a b_j **(-diag⟨1 / d_{jj}, 0⟩ $\bar{D}$ + I)** - distributed discrete sample is drawn for the next state,
   iii. in case the next state is an absorbing state then goes to 3 otherwise stay at 2 and repeat

   for j=1, 2....*n*;
3) do x_{ph} +=Erl(f_j, -d_{jj});
   done
4) Return x_{ph}.

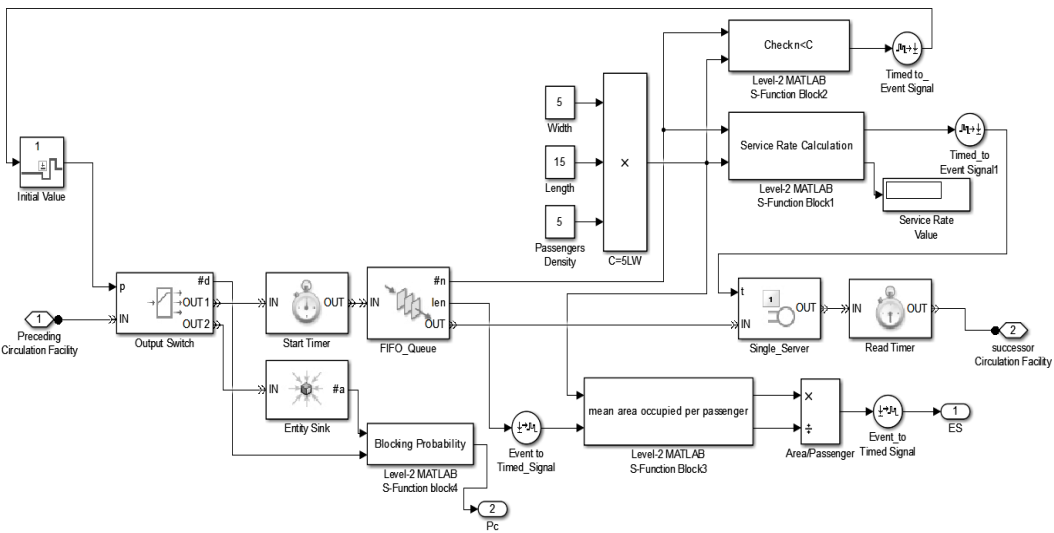
where, U is the uniformly distributed random number [0, 1]. Let b_j represent the row vector with 1 at position j. The 'Count Procedure' of drawing a sample from the Erlang distribution of length f_j is more efficient than drawing samples from the exponential distribution. The Erlang distribution requires a single logarithm operation opposed to the f_j logarithms when drawing individual exponential samples. This procedure instead of drawing exponential samples for each visit to a state j counts the number of visits and then draws one Erlang-distributed sample for each state.

6.2. PH-based DES Model Architecture

A PH-based DES model of the urban rail transit station circulation facilities is built in *SimEvents*® in this subsection, as shown in Figure 3a. In contrast to the PH-based analytical queuing model proposed in Hu et al 2015, the PH-based DES model eliminates the need to solve large matrix equations to estimate the performance measures.



(a) *SimEvents*® implementation of urban rail transit station circulation facilities network



(b) State-dependent service phase of the urban rail transit station circulation facility

Fig. 3**.** PH-based DES Model Architecture

Each circulation facility in PH-based DES model is described as a PH/PH (*n*)/C/C queueing system. An overview of the PH-based DES model of each circulation facility is presented in Figure 3b. The key components used in a PH-based DES model are as follows:

- The *Time-Based Entity Generation* block represents the source of passengers
- The *FIFO_Queue* block represents the passenger queueing space
- The *Single-Server* block stores the entities for certain period of time (entities in our case are passengers)
- The *Start and Read Timers* blocks report the time associated with the passengers
- The*Level-2 MATLAB® S-function* blocks compute and update different parameters in the PH-based DES model
- The *Event to Timed Signal* block and *Timed to Event Signal* block convert event-based signals to time-based signals and vice versa
- The *Constant* blocks are used to input different constant parameters values in the DES model while
- The *Display* blocks show the performance measure (output)
- The *SimOut* blocks export the values of performance measures from *SimEvents®* simulation to *MATLAB®* programming environment.

6.3. Passengers Generation Phase

In the PH-based DES model as shown in Figure 3a, the passengers are first generated at the entrance of the stairs on the platforms (No. 3, 9, 15 and 21) (see Figure 1) after alighting the train on Line 1 and Line 2 at the transfer station. The PH random variates are programmed in *Level-2 MATLAB® S*-function blocks (designated as S4) at passenger generation phase using 'Count Procedure' as discussed above. The two input parameters for the computation of PH random variates are the initial probability vector and the transient generator matrix that can be obtained by passenger arrival rate λ_a and SCV of arrival interval c_a^2 by using Equation (3), (4) and (5) respectively.

6.4. State-dependent Service Phase

After the generation, passengers will move forward to the circulation facilities. If the number of the passengers in the targeted facility is smaller than its capacity, passengers arriving at the i^{th} circulation facility form a queue and have to wait to be served. To implement this condition, the generated passengers are stored in the *FIFO_Queue* block before being delayed by the *Single_Server* block. After being served, the passengers will be sent to the successor circulation facility. During this process, they reduce the free spaces in the circulation facility and affect the walking speed of other passengers crossing the facility.

If the number of the passengers in the targeted facility has reached its capacity $C_i = 5L_iW_i$, the newly arrived passengers cannot enter the facility. To guarantee the number of passengers that enter the circulation facility do not overcome its overall capacity $C_i = 5L_iW_i$, the *Output Switch* is used to introduce another route for the passengers who cannot enter the circulation facility. When the successor circulation facility is not full, passengers will come out of it from the 1st entity port (OUT1), otherwise, passengers will come out from the 2nd entity port (OUT2).

Four *Level-2 MATLAB® S-function* blocks are used in this phase to calculate the state-dependent service time based on PH random variates, mean areas occupied per passengers 'ES_i', blocking probabilities $P_{c,i}$, and judging the number of passengers to prevent them from entry when maximum capacity $C_i = 5L_iW_i$ is reached as shown in Figure 4b. The state-dependent service time calculation depends on congestion in the circulation facility area. The capacity $C_i = 5L_iW_i$ and number of passengers (*n*) from the *FIFO_Queue* block are the input parameters of the *Level-2 MATLAB® S-function* blocks. They are used to compute the state-dependent service rate $\mu_{i,n}$ and SCV of state-dependent service rate $c_{i,s,n}^2$ using Equation (10) and Equation (11) respectively .Then the random number for service time will be generated in the same way used when generating arrival intervals.

The service time calculation block dynamically updates the service rates as a function of the number of passengers (*n*) for each circulation facility. At the

same time, two important performance measures are collected. The mean area occupied per passenger 'ES_i' is calculated by using area of each circulation facility $A_i = L_i W_i$ divided by mean queue length (len) obtained by the *FIFO_Queue* block. The blocking probability $P_{c,i}$ is calculated by using the number of passengers departed via the 2nd entity port of *Output Switch* divided by the total number of passengers departed via both 1st (OUT 1) and 2nd ports (OUT 2).

Before we develop the *simulation-optimization* approach, it is necessary to verify the accuracy of the proposed PH-DES model. Currently, no PH-based analytical model for the network is available. As it is proved in Hu et al 2015 that the M/G(*n*)/C/C model (Cruz et al. 2005) is a special case of PH-based queuing model and the PH-based queuing model can be converted into the M/G(*n*)/C/C model if c_a^2 and c_s^2 are equal to 1. Therefore, the existing M/G(*n*)/C/C network model is applied as a standard for the comparison.

A simple network constituting three corridors, each with size 8×2.5 m^2 in series, splitting and merging network topologies are analyzed. The passenger arrival rate is λ_a = 3ped/s in both approach. To compare on the same benchmark, the c_a^2 and c_s^2 are equal to 1 in the PH-based DES model and the c_s^2 in the M/G(*n*)/C/C model is also 1.Other parameters are the same in the two methods. The performance measures, including the mean number of passengers E[N], mean waiting time in queue E[W], blocking probabilities P_c and throughput θ are computed by the two methods. The results of PH-based DES model are obtained after 10 repetitions (each simulation last 20,000 units to make sure that the performance measures become stable). The results of the two methods are presented in Table 1. The comparison in Table 1 shows that PH-based DES Model has a smaller average relative error and indicates that PH-based DES model can be used with good accuracy in performance evaluation of urban rail transit stations circulation facilities.

Table 1. Comparison of PH-based DES Model and Analytical Model

	Corridor 1		Corridor 2		Corridor 3		Mean Relative Error (%)
	Analytical M/G(*n*)/C/C	PH-based DES	Analytical M/G(n)/C/C	PH-based DES	Analytical M/G(*n*)/C/C	PH-based DES	
Series Topology							
P_C	0.33	0.32	0.00	0.00	0.00	0.00	1.01
θ	2.01	2.00	2.01	2.00	2.01	2.01	1.01
E[N]	96.96	96.04	14.56	16.02	14.56	15.94	6.24
E[W]	48.31	47.95	7.26	8.1	7.26	8.02	6.86
Merging Topology							
P_C	0.33	0.32	0.33	0.32	0.53	0.52	2.72
θ	2.00	1.98	2.00	1.98	2.00	1.99	0.84
E[N]	99.51	98.41	99.51	98.33	99.76	98.59	1.17
E[W]	47.82	47.61	47.82	47.33	50.54	50.11	0.77
Splitting Topology							
P_C	0.33	0.32	0.00	0.00	0.00	0.00	1.01
θ	2.01	2.00	1.04	1.04	1.04	1.04	0.17
E[N]	96.96	95.41	7.75	7.70	7.75	7.70	0.97
E[W]	48.31	47.95	7.53	7.45	7.53	7.45	0.96

7. PH-based Simulation-Optimization approach for the widths design

Based on the PH-based DES model, we develop the PH-based *simulation-optimization* approach for the urban rail transit stations circulation facilities width design. The GA is used as an optimization approach in conjunction with the PH-DES model to determine the optimal widths of circulation facilities. The GA is implemented in the *MATLAB*® programming environment. The proposed PH-based *simulation-optimization* approach blends both the PH-based DES and GA to work together concurrently and find the optimal widths of the circulation facilities is presented below and the flow chart is presented in Figure 4.

- A set of N number of widths of circulation facilities $W_i = \{w_1, w_2, ...w_N\}$ to be optimized under the LOS 'B' and the blocking probability P_c below $p = 0.001$.
- The width set W has a domain set $D = \{d_1, d_2, ...d_N\}$.
- The multidimensional search space U (one for each width) is defined by

$$U = \{u = \{s_1, ...s_N\} \mid s_i \in d_i\}$$

- According to the TCQSM, the LOS of the circulation facilities is reflected by the mean area occupied per passenger ES_i, which means that the ES_i for the circulation facilities for a given LOS must fall within the range $[LOS_{LB,i}, LOS_{UB,i}]$, where $LOS_{LB,i}$ and $LOS_{UB,i}$ are the lower and upper bounds of the mean area occupied per passenger for the given LOS.
- The performance measures (outputs) $ES_i = \{ES_1, ...ES_N\}$ and $P_{c,i} = \{P_{c,1}, ...P_{c,N}\}$ are estimated by running the PH-based DES model of urban rail transit station circulation facilities (see Figure 4b).
- The mean area occupied per passenger *ES* for the circulation facilities is $ES_i = L_i W_i / n_i$, $(i = 1, 2.....N)$ from which we can see that mean area occupied per passenger will vary with the width W_i. Therefore, the mean area occupied per passenger for the circulation facilities can be expressed as a function of W_i, that is, $ES(W)_i = L_i W_i / n_i$

Therefore, the width optimization problem is to find the smallest widths that make sure that the Mean area occupied per passenger ES_i fall within the range $[LOS_{LB,i}, LOS_{UB,i}]$ and the blocking probability is smaller than the required value *p*, that is:

$$\min W_i$$
$$\text{s.t. } LOS_{LB,i} \le ES(W_i) \le LOS_{UB,i}$$
$$P_{c,i}(W) \le p$$

In this research, the *MATLAB*® GA toolbox released by The MathWorks™ is used. The default *MATLAB*® GA parameter settings are used, except for a decreased population size of 20 and an adjusted termination criterion if the weighted mean change in the fitness function value over x generations is less than 0.01, the algorithm stops.

The GA parameters and their values are listed below. A description and lists of possible values as well as the algorithm description can be found in The MathWork™.

Population
- Population Size: 20
- Creation Function: Uniform
- Initial Population: []
- Initial Score: []

Reproduction
- Elite Count: 2
- Crossover Fraction: 0.8

Mutation
- Mutation Probability: 0.01

Termination Criteria
- Function Tolerance: 0.01
- Stall Generation: 10
- Time limit: Inf

It should be noted that the population size, stall generation and the termination criteria are adapted for this study. It is possible that changes of other parameters would lead to better optimization results but in this research we develop an integrated PH-based DES model with GA and assess the comparison of width obtained by using this PH-based simulating-optimization and other existing model such as M/G(*n*)/C/C and D/D/1/C, therefore further experiments with different parameters are not undertaken in the scope of this research.

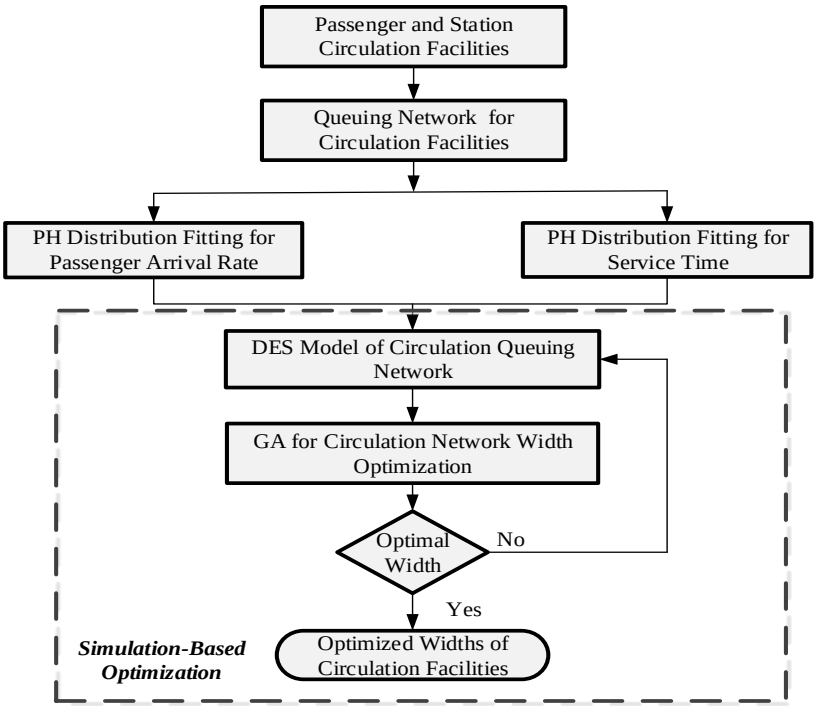


Fig. 4. PH-based *Similation-Optimization* for the circulation facilities widths design

According to the optimization model, simulation-based optimization approach is proposed. The *MATLAB*® programming environment is used to run the PH-based DES model by using *'sim'* command. Since *MATLAB*® offers parallel DES and optimization, therefore the performance measure values from the PH-based DES model are transferred from *SimEvents*® to *MATLAB*® environment by using *'yout'* block. If the constraint function is not satisfied, the GA set new values of parameter to be optimized by using *'set_param'* command and the loop continues until the optimal results are obtained or termination criteria satisfy.

The *simulation-based optimization* works as follow: At first, the interval containing the upper and lower bounds of circulation facilities width $U=[W_{UB,i},W_{LB,i}]$ is defined which is supposed to contain the optimal width $W_{opt,i}$ of the circulation facilities. The PH-based DES model runs initiate with an arbitrary value from the defined interval to simulate the performance of the circulation facilities and obtain the performance measure (ES_i and $P_{c,i}$) when simulation system reaches the steady state condition (when the performance measures become stable). Then the Genetic Algorithm (GA) that is programmed in *MATLAB*® adjusts the widths W_i according to the value of ES_i and $P_{c,i}$ until the optimal widths $W_{opt,i}$ are found.

To improve the efficiency of the optimization model, a function tolerance η is defined. If the relative change in the objective is less than or equal to the η then the corresponding W_i can be approximately considered as the optimal width $W_{opt,i}$. If the difference is larger than η , the GA will replace W_i from the defined interval $[W_{UB,i},W_{LB,i}]$ and set the new width value W_i in the PH-based DES model for next iteration to obtain the ES_i and $P_{c,i}$ by the same means. The iterations continue until the relative change in the best fitness

function value is less than or equal to η and the corresponding width W_i is the optimal width $W_{opt,i}$. The minimum allowable width under the TCQSM is 1 meter. But instead of using 1meter as lower bound of width, we set the upper and lower bound calculated. The width design under the LOS 'B' will fall in this range and less likely to trap in the local optimum. It should be noted that the search space obtained by using the min and max values of ES neglected the randomness and state-dependent. The search space is used to find the optimal result is obtained from TCQSM. Moreover, after reviewing several literatures, one of the main reasons to use GA is that it searches dozens or hundreds of parts of the search space simultaneously which means that it is less likely to become stuck in "local minima" as the others traditional optimization approaches quite often do. The more details regarding the upper and lower bounds of width with an example to make it clearer is presented. The Exhibit 7-3 (Pedestrian Level of Service in walkways) of TCQSM presents the upper and lower bound values of flow per unit width (ped/m/min) under the different LOS. We use these values as our benchmark to define the upper and lower limit of width. An example is presented below.

Let us consider we design under LOS 'B' for peak-hour factor of 0.3 and we have an hourly volume given as 5000 ped/h. The upper and lower limit of flow per unit width under LOS 'B' is 33 and 23, respectively from Exhibit 7-3. According to TCQSM, the width of the walkway can be obtained as:

Upper bound of width

$$W_{UB} = \frac{5000}{(0.3)(60)(23)} = 12m$$

$$W_{LB} = \frac{5000}{(0.3)(60)(33)} = 8.4m$$

$[W_{UB} + 2, W_{LB}] = [14, 8.4]$ is used as the upper and lower limit under this condition. The upper bound is increased by an increment of 2 as it is expected to have a design width higher than upper bound due to increase in SCV of arrival internal. It should be noted that these upper and lower bound values are estimated by using TCQSM width design procedure that neglects randomness and state-dependence. We use these values only to define our search space and initiate our PH-based DES model run.

8. Computational Experiments

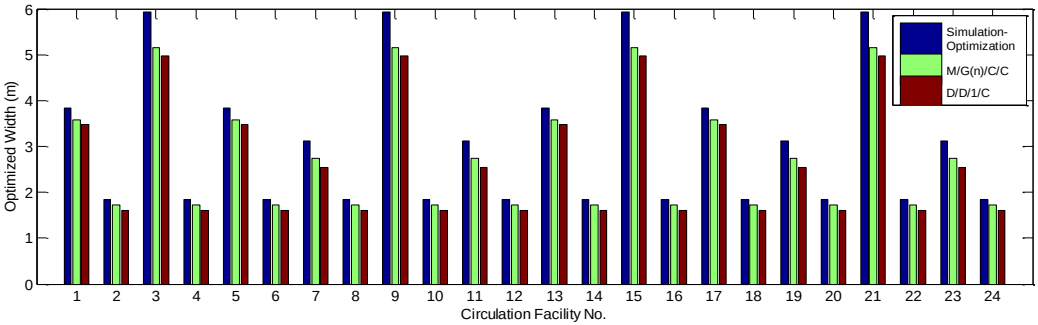
In the following section, we will use the proposed PH-based *simulation- optimization* approach to optimally design widths of circulation facilities in the urban rail transit stations. We will design the width for the circulation facilities in Figure 1. The required input parameters for the width design such as passenger arrival rate, SCV of arrival interval, three representative points for walking speed in corridors and stairs and the lengths of circulation facilities are predetermined. The passenger arrival rates $\lambda_i \ (i = 3,9,15,21)$ values are 2ped/s and 3ped/s according to its actual range in the urban rail transit station. It can also be calculated by using Equation (3). Similarly, the SCV of arrival interval $c_{i,a}^2 \ (i = 3,9,15,21)$ values are 100, 300 and 500 according to its actual range in the urban rail transit stations. The SCV of arrival interval can also be determined by using Equation (4) and (5). The three representative points for walking speed in the corridor circulation facilities are $(v_{i,1} = 1.50, \delta_{i,1} = 0.50)$, $(v_{i,a} = 0.64, \delta_{i,a} = 0.21)$ and $(v_{i,b} = 0.25, \delta_{i,b} = 0.08)$ respectively (Hu et al. 2015), while, the three representative point of walking speed in the stairs facilities are $(v_{i,1} = 0.75, \delta_{i,1} = 0.25)$, $(v_{i,a} = 0.32, \delta_{i,a} = 0.11)$ and $(v_{i,b} = 0.12, \delta_{i,b} = 0.04)$ respectively. The state-dependent SCV ($c_{i,s,n}^2$) of service time of the i^{th} circulation facility can be calculated by using Equation.(11). The lengths of corridor facilities are 10m while the lengths of stairs facilities are 15 m. The design widths of all circulation facilities are obtained under the LOS 'B' i.e., $ES_i \geq 1.4$ m^2/ped and $ES_i \geq 2.3$ m^2/ped for stairs and corridors, respectively. The blocking probability $P_{c,i}$ should be below 0.001.

The widths designed by the proposed method are compared with the widths obtained by the existing M/G(*n*)/C/C (Cruz et al., 2005) and D/D/1/C (Kittelson et al., 2003) analytical approaches. The

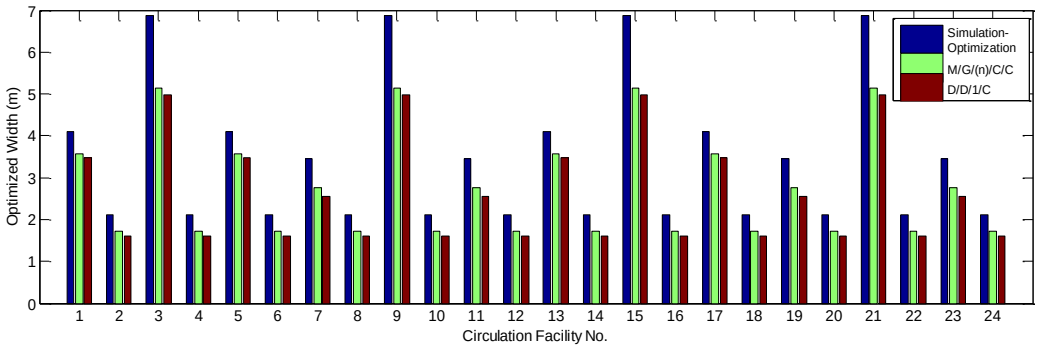
design procedure in TCQSM is similar to uses a fixed arrival rate and a fixed service time, which is essentially a D/D/1/C analytical queuing model (Jiang et al 2015). The SCV is equal to 1/30 (0.03) as it neglect randomness and state-dependence. We use the D/D/1/C queuing model to represent the width design procedure of TCQSM for the circulation facilities.

The widths designed by the three methods, the proposed PH-based *simulation-optimization* approach, the M/G(*n*)/C/C analytical model, and the D/D/1/C analytical model, are presented in Figure 5 and 6. The figures reveal some important and interesting findings:

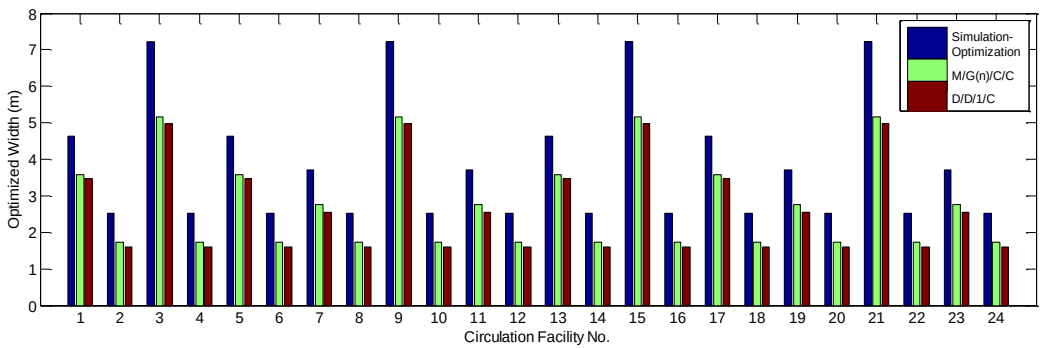
1) The design widths obtained by the PH-based *simulation-optimization* are greater than that of the M/G (*n*)/C/C and the D/D/1/C analytical model for all the arrival rates and SCV of arrival intervals. Figure 5 illustrates the design widths of the three approaches sunder the same arrival rate $\lambda = 2$ ped/s and different SCV of arrival interval (100, 300 and 500). Compared to the width designed by the D/D/1/C model, the average increase in the width of the proposed method is 0.43 m when the SCV of arrival intervals is 100, 0.83 m when the SCV of arrival intervals is 300, and 1.21 m when the SCV of arrival intervals is 500. Compared to the width designed by the M/G (*n*)/C/C model, the average increase in the width for the *simulation-optimization* is 0.29 m when the SCV of arrival intervals is 100, 0.69 m when the SCV of arrival intervals is 300, and 1.08 m when the SCV of arrival intervals is 500. Figure 6 shows similar trend for arrival rate $\lambda = 3$ ped/s. We can see the circulation facilities designed by the PH-based *simulation-optimization* approach has larger widths because it describes the circulation system more elaborate and considers both the LOS and the blocking probability.
2) The widths of the M/G (*n*)/C/C and D/D/1/C stay the same when the SCV of arrival interval changes from 100 to 500. On the contrary, the widths for the PH-based *simulation-optimization* increase with the increase in the SCV of arrival interval. This is because in the D/D/1/C, the randomness and state-dependence are completely ignored while in M/G(*n*)/C/C the passenger flow is assumed as a free flow where the SCV of arrival interval equals 1. Therefore, the design width of the two methods will not increase with the SCV of arrival interval. This result shows that the design methods based on the M/G (*n*)/C/C and D/D/1/C models are not applicable in practical systems where the SCV of arrival interval is far more than 1. On the contrary, the width of the proposed PH-based *simulation- optimization* approach is sensitive to the SCV of arrival interval.
3) For all the three design approaches, the widths of circulation facilities increase with the increase in passenger arrival rate when the SCV of arrival interval remains same. It is expected because of the fact that these design approaches are sensitive to the arrival rate. When the arrival rate increase from 2 to 3 ped/s, the average increase of the PH-based *simulation-optimization*, M/G (*n*)/C/C and D/D/1/Care 1.51 m(47%), 0.82 m(31%) and 0.76 m(31%) respectively. The PH-based *simulation-optimization* approach has a larger growth than the other two methods.
4) For all arrival rates and SCV of arrival intervals, the design widths of stairs facilities are greater than corridors facilities. It is quite obvious because of the fact that passengers' walking speed the on stairs is slower than that in the corridors. Thus more passengers are stranded in the stairs facilities, which will cause blocking and reduction in the mean area occupied per passenger 'ES_i'. Therefore, stairs require more width to keep the mean area occupied per passenger in the LOS 'B' range and blocking probability below 0.001. In addition, the widths of corridors No. 1,5,13 and 17 are greater than the other corridors because of merging topologies that require more widths to keep the '*ES*' above 2.3m^2/ped and blocking probability below 0.001.
5) It is observed that the average difference in the design widths of M/G (*n*)/C/C and D/D/1/C is 0.14 m, which is much smaller than the difference between the width of the M/G (*n*)/C/C or D/D/1/C model with the width of the PH-based *simulation-optimization*. This also illustrates that proposed approach can reveal the extra requirement on width which is ignored by the existing methods.



(a) Comparison of design widths for passenger arrival rate $\lambda = 2$ ped/s and SCV=100

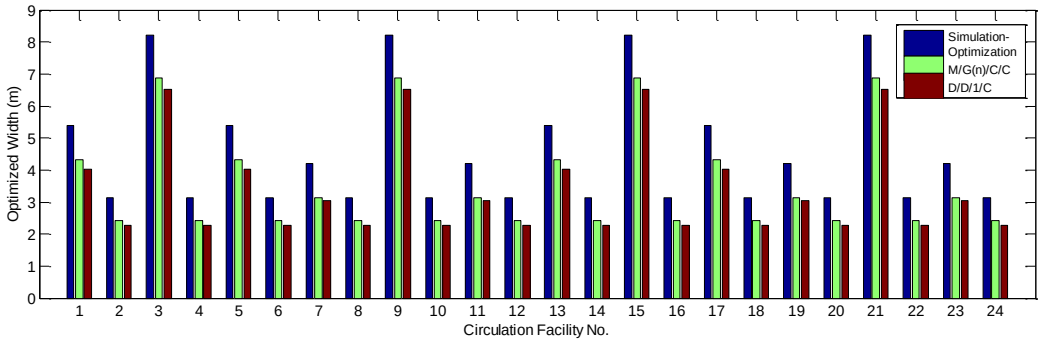


(b) Comparison of design widths for passenger arrival rate $\lambda = 2$ ped/s and SCV=300

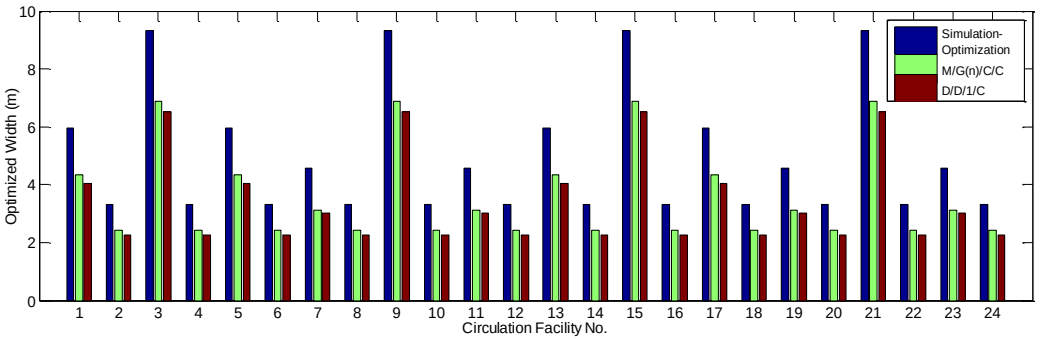


(c) Comparison of design widths for passenger arrival rate $\lambda = 2$ ped/s and SCV=500

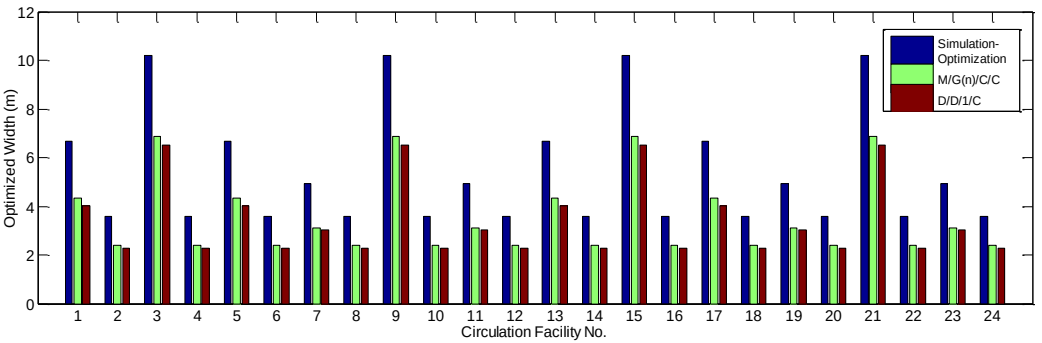
Fig. 5. Design widths comparison for passenger arrival rate $\lambda = 2$ ped/s



(a) Comparison of design widths for passenger arrival rate $\lambda =$ 3ped/s and SCV=100



(b) Comparison of design widths for passenger arrival rate $\lambda = 3$ ped/s and SCV=300



(c) Comparison of design widths for passenger arrival rate $\lambda = 3$ ped/s and SCV=500

Fig. 6. Design widths comparison for passenger arrival Rate $\lambda = 3$ ped/s

9. Conclusions and Future work

This paper proposes a PH-based *simulation-optimization* approach by integrating a PH-based DES model and GA for the widths design of circulation facilities in urban rail transit station. The proposed approach overcomes the shortcomings in the existing design approaches by fully consider the randomness and state dependence in the PH-based DES model and consider the requirement on both LOS and blocking probability in the optimization. A

comparison is made between the M/G (*n*)/C/C model and the proposed PH-based DES model to verify the accuracy of the latter one. The results show that the PH-based DES model has achieved clear consistency with the analytical approach. In addition, the experiments on width design are carried out by comparing the PH-based *simulation-optimization* approach with the existing design approaches.

The numerical experiments reveal some interesting findings: (1) The circulation facilities designed by the PH-based *simulation-optimization* approach has larger widths compared with that designed by the existing methods; (2) The width of the proposed method increase with the SCV of arrival interval, while the widths of the design methods based on the M/G (*n*)/C/C and D/D/1/C models stays the same where the SCV of arrival interval increases; (3) The width of the proposed method increase faster than the other two methods when the arrival rate increases; (4) Under the same passenger flow conditions, stairs require more width to meet the requirement on LOS and blocking probability.

This new proposed PH-based *simulation-optimization* approach, integrating PH-based DES and optimization can help the planners and designers of urban rail transit station to make decisions regarding urban rail transit station design. This approach can also be applied to design circulation facilities in other public buildings such as shopping malls and hospitals etc., if the pedestrian peak hour flow, circulation facilities lengths, the desired LOS and peak hour factors are known. The PH-based *simulation-optimization* is particularly useful in situations where the analytical expressions are too complex to obtain. At the same time, this approach can serve as an important tool for verifying the PH-based analytical model developed in Hu et al. (2015).

This paper only considers rectangular circulation facilities for evaluation and design purpose. Other complicated circulation facilities that are not rectangular can be divided into several rectangular facilities and then be evaluated in the same way. The principle procedure of circulation facilities transformation into a single server queuing system remains the same. In addition, we only consider the unidirectional passenger flow in this paper. But the model can also deal with bidirectional or multidirectional passengers flow by only adjusting the speed parameters. Moreover, the queuing system is considered to be a loss queue without feedback. However, feedback always exists in circulation facilities when congestion happens. A PH-based DES model for a feedback queuing system will be addressed in our future research.

Acknowledgment

We would like to express our sincere acknowledgment to National Natural Science Foundation of China (Serial No. 51578465 and 71402149), Basic Research Project of Sichuan Province, the Chinese government for funding of PhD doctoral program at Southwest Jiaotong University and the colleagues of National United Engineering Laboratory of Integrated and Intelligent Transportation at Southwest Jiaotong University, Chengdu for their support and valuable advice.

References

[1] ALFA, A. S., and ZHAO, Y. Q., 2000. Overload analysis of the PH/PH/1/K queue and the queue of M/G/1/K type with very large K. *Asian Pacific Journal of Operational Research*, 17, 122-136.

[2] BANKS, J. (2010). *Discrete-event System Simulation*, Prentice Hall.

[3] CASSANDRAS, C. G., & LAFORTUNE, S., 2009. *Introduction to Discrete Event Systems*: Springer US.

[4] CHEAH, J., and SMITH, J. M., 1994. Generalized M/G/C/C state dependent queueing models and pedestrian traffic flows. *Queueing System*, 15(1-4), 365-386.

[5] CHEN, S. K., and S. Liu, XIAOM,X., 2012. M/G/C/C-based Model of Passenger Evacuation Capacity of Stairs and Corridors in Urban Rail Transit Station. *Journal of China Railway Society*, 34, 7-12.

[6] CRUZ, F. R. B., MACGREGOR SMITH, J., AND MEDEIROS, R. O., 2005. An M/G/C/C state-dependent network simulation model. *Computers & Operations Research*, 32(4), 919-941.

[7] CRUZ, F. R. B., SMITH, J. M., AND QUEIROZ, D. C., 2005. Service and capacity allocation in M/G/C/C state-dependent queueing networks. *Computers & Operations Research*, 32(6), 1545-1563.

[8] FIGUEIRA G, ALMADA-LOBO B., 2014. Hybrid simulation–optimization methods: A taxonomy and discussion[J]. *Simulation Modelling Practice and Theory*, 46, 118-134.

[9] FU, M. C., 2002. Feature Article: Optimization for simulation: Theory vs. Practice. *INFORMS Journal on Computing*, 14(3), 192-215.

[10] HAGENDORF, O., PAWLETTA, T. and LAREK, R., 2013. An approach for simulation-based parameter and structure optimization of MATLAB/Simulink models using evolutionary algorithms. *SIMULATION*.

[11] HASSANNAYEBI E, SAJEDINEJAD A, MARDANI S., 2014. Urban rail transit planning using a two-stage simulation-based optimization approach. *Simulation Modelling Practice and Theory*, 49, 151-166.

[12] HU, L., JIANG, Y., ZHU, J., and CHEN, Y., 2015. A PH/PH state-dependent queuing model for metro station corridor width design. *European Journal of Operational Research*, 240(1), 109-126.

[13] HUBSCHER-YOUNGER, T., MOSTERMAN, P. J., DELAND, S., ORQUEDA, O., and EASTMAN, D., 2012. Integrating discrete-event and time-based models with optimization for resource allocation. *Proc., Simulation Conference (WSC), Proceedings of the 2012 Winter*, 1-15.

[14] JACYNA, M., WASIAK, M., LEWCZUK, K., KŁODAWSKI, M., 2015. Simulation model of transport system of poland as a tool for developing sustainable transport. *The Archives of Transport*, 31(3), 23-35.

[15] JIANG, Y., and LIN, X., 2013. Simulation and Optimization of the Ticket Vending Machine Configuration in Metro Stations Based on Anylogic Software. *Fourth International Conference on Transportation Engineering*, 754-760.

[16] JIANG, Y., HU, L., and LU, G., 2010. Determined method of subway footway width based on queuing theory. *Journal of Traffic and Transportation Engineering*, 10, 61-67.

[17] JIANG, Y., HU, L., ZHU, J., and CHEN, Y., 2013. PH fitting of the arrival interval distribution of the passenger flow on urban rail transit stations. *Applied Mathematics and Computation*, 225, 158-170.

[18] JIANG, Y., ZHU, J., HU, L., LIN, X., and KHATTAK, A., 2016. A G/G(n)/C/C state-dependent simulation model for metro station corridor width design. *Journal of Advanced Transportation*, 50, 273–295.

[19] KAAKAI F, HAYAT S, EL MOUDNI A., 2007. A hybrid Petri nets-based simulation model for evaluating the design of railway transit stations. *Simulation Modelling Practice and Theory*, 15(8), 935-969.

[20] KHALID, R., M. NAWAWI, M. K., KAWSAR, L. A., GHANI, N. A., KAMIL, A. A., and MUSTAFA, A., 2013. A Discrete Event Simulation Model for Evaluating the Performances of an M/G/C/C State Dependent Queuing System. *PLoS ONE*, 8(4), e58402.

[21] KITTELSON, Associates, Administration, U. S. F. T., Program, T. C. R., Corporation, T. D., and Board, N. R. C. T. R., 2003. *Transit Capacity and Quality of Service Manual*, Transportation Research Board of the National Academies.

[22] KRISHNAMOORTHY, A., BABU, S., and NARAYANAN, V. C., 2008. MAP/(PH/PH)/c Queue with Self-Generation of Priorities and Non-Preemptive Service. *Stochastic Analysis and Applications*, 26(6), 1250-1266.

[23] LEWCZUK, K., 2015. The concept of genetic programming in organizing internal transport processes. *Archives of Transport*, 34(2), 61-74.

[24] LØVÅS, G. G., 1994. Modeling and simulation of pedestrian traffic flow. *Transportation Research Part B: Methodological*, *28*(6), 429-443.

[25] SMITH, J. M. G., 1991. State-dependent queueing models in emergency evacuation networks. *Transportation Research Part B: Methodological*, 25(6), 373-389.

[26] MESSAC, A., 2015. *Optimization in Practice with MATLAB*, Cambridge University Press.

[27] MITCHELL, D. H., and MACGREGOR SMITH, J., 2001. Topological network design of pedestrian networks. *Transportation Research Part B: Methodological*, 35(2), 107-135.

[28] MIYAZAWA, M., SAKUMA, Y., & YAMAGUCHI, S., 2007. Asymptotic Behaviors of the Loss Probability for a Finite Buffer Queue with QBD Structure. *Stochastic Models*, *23*(1), 79-95.

[29] NEUTS, M. F., 1981. *Matrix-geometric Solutions in Stochastic Models: An Algorithmic Approach*, Dover Publications.
[30] SADRE, R., 2007. *Decomposition-based analysis of queueing networks*. University of Twente, Twente.
[31] SADRE, R., and HAVERKORT, B., 2011. Decomposition-Based Queueing Network Analysis with FiFiQueues. *Queueing Networks*, R. J. Boucherie, and N. M. van Dijk, eds., Springer US, 643-699.
[32] SWISHER, J. R., HYDEN, P. D., JACOBSON, S. H., and SCHRUBEN, L. W., 2000. Simulation optimization: a survey of simulation optimization techniques and procedures. *Proceedings of the 32nd conference on Winter simulation*, Society for Computer Simulation International, Orlando, Florida, 119-128.
[33] TEKNOMO, K., 2006 Application of microscopic pedestrian simulation model. *Transportation Research Part F: Traffic Psychology and Behaviour*, 9(1), 15-27.
[34] TREGENZA, P., 1976. *The design of interior circulation*, Van Nostrand Reinhold.
[35] XU, X.., LIU, J., LI, H., AND HU, J., 2014. Analysis of subway station capacity with the use of queueing theory. *Transportation Research Part C: Emerging Technologies*, 38, 28-43.
[36] YUHASKI, S. J., and SMITH, J. M., 1989. Modeling circulation systems in buildings using state dependent queueing models. *Queueing Systems*, 4(4), 319-338.

ARCHIVES OF TRANSPORT www.czasopisma.pan.pl PAN www.journals.pan.pl ***ISSN (print):*** *0866-9546*
Volume 43, Issue 3, 2017 ***e-ISSN (online):*** *2300-8830*
DOI: *10.5604/01.3001.0010.4228*

NON-PARAMETRIC MACHINE LEARNING METHODS FOR EVALUATING THE EFFECTS OF TRAFFIC ACCIDENT DURATION ON FREEWAYS

Ying Lee[1], Chien-Hung Wei[2], Kai-Chon Chao[3]

[1] National Kaohsiung Marine University, Kaohsiung, Taiwan

[2] National Cheng Kung University, Tainan, Taiwan

[3] THI Consultants Incorporation, Taipei, Taiwan

[1]e-mail: yinglee1017@gmail.com
[2]e-mail: louiswei@mail.ncku.edu.tw
[3]e-mail: stone_bhm1990@hotmail.com

Abstract: *Traffic accidents usually cause congestion and increase travel-times. The cost of extra travel time and fuel consumption due to congestion is huge. Traffic operators and drivers expect an accurately forecasted accident duration to reduce uncertainty and to enable the implementation of appropriate strategies. This study demonstrates two non-parametric machine learning methods, namely the k-nearest neighbour method and artificial neural network method, to construct accident duration prediction models. The factors influencing the occurrence of accidents are numerous and complex. To capture this phenomenon and improve the performance of accident duration prediction, the models incorporated various data including accident characteristics, traffic data, illumination, weather conditions, and road geometry characteristics. All raw data are collected from two public agencies and were integrated and cross-checked. Before model development, a correlation analysis was performed to reduce the scale of interrelated features or variables. Based on the performance comparison results, an artificial neural network model can provide good and reasonable prediction for accident duration with mean absolute percentage error values less than 30%, which are better than the prediction results of a k-nearest neighbour model. Based on comparison results for circumstances, the Model which incorporated significant variables and employed the ANN method can provide a more accurate prediction of accident duration when the circumstances involved the day time or drunk driving than those that involved night time and did not involve drunk driving. Empirical evaluation results reveal that significant variables possess a major influence on accident duration prediction.*

Key words: *accident duration, correlation, artificial neural networks, k-nearest neighbour method.*

1. Introduction

Traffic accidents usually cause considerable speed reduction and congestion on freeways due to lane closures or obstacles. For fifty large U.S. urban areas, the cost of extra travel time and fuel consumption due to congestion annually amounts to approximately $37.5 billion (Winston & Langer, 2006). To mitigate the impacts due to congestion, traffic management centres usually develop accident management programs. The aims of these programs include exploration of the important factors of accidents, detection of accidents, and provision of accident information forecasts. The impacts of each accident, i.e., duration and resulting congestion queue, may be affected by different features. Relevant features include continuous and/or categorical data, such as accident type, accident characteristics, the number of injuries or fatalities, illumination, type of vehicle involved, road geometry characteristics, and weather conditions. If these data can be processed and analysed effectively, traffic patterns under the influence of accidents could be adequately characterized for various applications in transportation.

Therefore, the objectives of this study are to collect, cross-check, and integrate accident features and traffic data for an accident duration forecasting model on a freeway. The model is based on a related accident database maintained by several public agencies. The proposed forecasting models apply a correlation analysis to select significant variables and employ k-nearest neighbour (kNN) and artificial neural network (ANN) approaches to develop a relationship between the selected variables and

accident duration. To demonstrate the performance of the proposed procedure, model performance evaluation is conducted to compare the prediction performance of models with and without correlation analysis and compare the prediction performances for the significant circumstances.

The remainder of this paper is divided into the following sections. Relevant literature is reviewed and assessed in Section 2. Section 3 presents data sources and data analysis. Section 4 provides a brief introduction to the methodologies and a model evaluation indicator. Section 5 illustrates the evaluation results of four prediction models. Finally, Section 6 presents concluding remarks and suggestions for future research.

2. Literature Review

2.1. Accident duration forecast

Many types of incidents occur on highways. Whether it is a serious traffic accident or a falling object, the event can be referred to as an incident that occurs on the road. To reduce the uncertainty of travellers during an incident, several researchers have investigated the relationship between incident duration and traffic/incident data to estimate/forecast accident duration.

Kim and Chang (2011) developed a hybrid prediction model for freeway incident duration. It consists of a rule-based tree model (RBTM), a multinomial logit model (MNL), and a naïve Bayesian classifier (NBC). The decision tree model involves a five-step procedure. It classifies the incident duration data from a database according to incident type, and constructs a rule-based tree under the incident conditions. The results show that incident durations of 120 to 180 minutes and 180 to 240 minutes have satisfactory outcomes. The model performs well for incidents of less than 60 minutes or longer than 300 minutes.

Zhan et al. (2011) applied a regression method and the M5P tree algorithm to predict the lane clearance time of an incident for five scenarios. The model inputs included time of day, day of the week, lighting condition, the number of vehicles involved in the incident, vehicle type involved in the incident, and the number of lanes occupied in the incident. The results of the model showed that incidents that occurred during weekends or those that involved buses or trucks have longer lane clearance times. When the incidents occurred during the daytime period on weekdays, the lane clearance times were shorter. The mean absolute percent error (MAPE) of model performance during prediction was about 42%. When the incident duration was longer than 30 minutes, the prediction error increased and the MAPE value was higher than 78%.

Khattak et al. (2012) analysed traffic incidents and presented iMiT (incident management integration tool) to dynamically predict incident durations. Based on a statistical regression method, the prediction model incorporated time of day, weather conditions, incident location, the number of vehicles involved in the incident, and incident type as the inputs. The MAPE of model performance in estimation and prediction was lower than 55% and displayed reasonable estimation and prediction results.

Li (2015) applied a survival analysis model to develop an incident duration prediction model during three incident duration stages. When the incident duration was between 15 and 60 minutes, the MAPE of model performance was lower than 47% and exhibited reasonable prediction behaviour. When the incident duration was short (less than 15 minutes) or long (greater than 60 minutes), the prediction error was large and the MAPE value was higher than 61%.

Chung et al. (2015) proposed an accelerated failure time model to forecast accident duration and evaluate model performance for the number of lanes blocked. Their results indicated that the accident duration with no blocked lanes was less than those with two or three blocked lanes. However, the accident duration with one blocked lane was less than those with no blocked lanes.

Most studies agree that the data or information collected from management processes can improve the accuracy of predicted incident duration for model development. For incident duration model development, Qi and Teng (2008) defined four categories of input variables according to a USA incident database. Variables used in their model included:

- Weather characteristics: sunny, rainy, and snowy
- Temporal characteristics: AM peak, PM peak, night, and weekday
- Incident characteristics: lanes, property, severity, debris, road repair, and pothole
- Involved vehicle characteristics: bus, van, and truck

During the past few years, a variety of methods have been applied to develop freeway accident duration estimating/forecasting models. The most representative approaches can be classified into the following categories: multivariate regression (Garib et al., 1997; Smith K. & Smith B., 2001; Valenti et al., 2010), fuzzy logic model (Choi, 1996; Dimitriou & Vlahogianni, 2015), artificial neural network (Wang et al., 2005), and survival (Chung et al., 2015; Nam & Mennering, 2000; Chung, 2010; Hojati et al., 2013). Representative studies on highway accident duration prediction over the decade are summarized in Table 1. After assessing the methods frequently employed in the literature, survival analysis (accelerated failure time model) is a popular approach for most researchers and demonstrates acceptable results in freeway accident duration estimation or prediction.

This research differs from most previous studies, which used a regression method or an accelerated failure time model as the key analytical technique for model development. Two non-parametric machine learning methods, namely kNN and ANN, are demonstrated in the freeway accident duration prediction models and performance assessment in this study. Both methods are suitable for modelling complex systems and often achieve a reliable performance. Many studies have demonstrated that kNN and ANN have the potential to accurately predict traffic conditions on highways (Chien et al., 2002; Vlahogianni & Karlaftis, 2013) or on other traffic issue (Spławińska, 2015; Pamula, 2012). Thus, kNN and ANN were chosen as the key analytical techniques in this study.

2.2. Feature selection with correlation analysis

Most researches incorporate high-dimensional data to describe and distinguish complex objects. However, large feature vectors may result in some disadvantages to the model, such as longer model training time and more noise in model development. To avoid these problems, the feature vectors must be properly reduced (Lee & Wei, 2009).

Correlation is a technique for determining whether a linear relationship exists between two variables. The closer the correlation coefficient is to ±1, the stronger the linear relationship between the two variables is. Therefore, conducting a correlation analysis is useful for distinguishing significant independent features from dependent features before model development.

Zhang (2000) presented a prediction algorithm using artificial neural networks. The model was determined by correlation analysis. The parameters of the model can be obtained through nonlinear optimization. Preliminary studies showed that this approach can yield reasonably accurate results.

Table 1. Recent studies of highway accident duration prediction

Researcher	Methodology	Characteristics for model input	Best model performance	Study area
Chung, 2010	Accelerated failure time model	Temporal, Involved vehicle, Accident	MAPE<47%	Korea
Zhan et al., 2011	Regression	Temporal, Involved vehicle, Accident	MAPE<42.7% RMSE<63.46	USA
Khattak et al., 2012	Regression	Temporal, Weather, Accident, Location	MAPE<218% RMSE<17.47	USA
Hojati et al., 2013	Accelerated failure time model	Temporal, Weather, Accident, Traffic		
Li, 2015	Accelerated failure time model	Accident, Season	MAPE<238% RMSE<39.06	China
Li et al., 2015	Competing risk mixture model	Temporal, Traffic, Vehicle, Location	MAPE<94.7% RMSE<26.61	Singapore
Dimitriou and Vlahogianni, 2015	Fuzzy	Weather, Accident, Traffic	MAPE<36%	-
Chung et al., 2015	Accelerated failure time model	Temporal, Weather, Accident, Traffic	-	Taiwan

Guo and Nixon (2009) applied a correlation method to select the important features as the inputs for a pattern recognition model. The experimental results showed that the model selected 37 features from 73 features by the correlation method and achieved 90% classification accuracy rate in pattern recognition.

Based on the discussed literatures, it is clear that conducting research on accident duration is as important as on travel time prediction during an incident. In order to reduce the impact of incidents on travel time prediction, this study identifies significant accident features and develops accident duration prediction models.

3. Data

3.1. Study site

This study selected the Taiwan National Freeway No. 5 (from the Nan-Gang system interchange to the Su-Ao interchange) as the site of the case study. This double-lane road is 54-kilometers long and has seven interchanges as indicated in Figure 1. The distance between two neighbouring vehicle detectors (VDs) is about 2 km. There are five tunnels in the case study site, including the Hsueh-Shan Tunnel, which is the fifth longest tunnel in the world.

3.2. Data sources

Currently, two public agencies, namely the National Police Agency (NPA) and the National Freeway Bureau (NFB), maintain separate raw data regarding traffic accident information on National Freeway No. 5 in Taiwan. The traffic data from NFB includes incident duration and location. The accident data from NPA is the primary source providing detailed information of accident features and environmental factors at an accident site, such as the number of fatalities/injuries, weather conditions, and pavement conditions. To incorporate all information in this study, all data from these two databases require integration and cross-checking. The relevant features of these two databases are listed below:

The National Freeway Bureau

- Incident duration: response time and clearance time;
- Direction: north or south;
- Location: the mileage on National Freeway No. 5;
- Information of involved vehicle: name and phone number of driver, number of vehicle;
- The status of towing: towing or not, including leaving the vehicle on its own and clearing

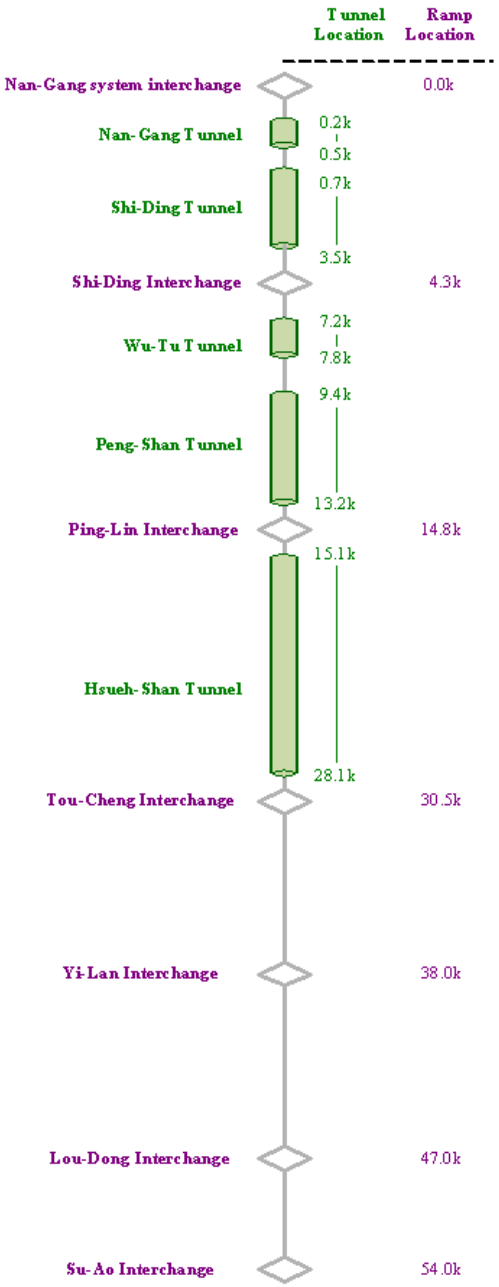


Fig. 1. Layout of Freeway No. 5

The National Police Agency
- Time: possible time of occurrence of confirmed accidents, response time;
- Location: the mileage on National Freeway No. 5;
- Direction: north or south;
- Injuries: number of injuries;
- Weather condition: sunny, rainy, or stormy;
- Type of road: tunnel or elevated road;
- Lighting condition: day time (exclude periods of dawn and dusk) or night time (includes tunnels and underpasses);
- Involved vehicles: type of vehicle involved (e.g., small truck, bus, tractor-semi trailer);
- Accident severity: A1, A2, or A3;
- Pavement condition: dry or wet;

In order to obtain the information of accident duration, this study integrates the above database with the Tow Truck Service Report database maintained by the NFB. Thus, the start time and clear-up time of an accident is practically available to evaluate the associated impact, i.e., accident duration.

A total of 239 accidents on National Freeway No. 5 were recorded during 2012 in the NPA database. However, most accident records are for the purpose of liability appraisal rather than for accident duration prediction. Therefore, integration with the NFB database is required due to the availability of clearance time of accidents. Consequently, a data set of 49 accidents on National Freeway No. 5 is obtained that combines the response time (Nation Police Agency) and clearance time (National Freeway Bureau).

The NFB installs VDs on highways to record traffic data such as speed, volume, and occupancy. Traffic patterns and variations during an accident can be adequately characterized by these data. Therefore, the accident duration can be obtained and verified by analysing these data. This study also incorporates the average speed and average volume as the model features. The traffic data from the VD were accumulated at an interval of five minutes.

3.3. Accident duration

The accident duration in this study represents the period between the time an accident is reported and the time when all handlers leave the accident site. The minimum, maximum, and average durations for 46 accidents were 14, 108, and 42 minutes. Table 2 shows the relative frequency of durations for 46 accidents. For about 56.5% of the accidents, the duration was less than 39 minutes. The percentage of accidents with durations between 40 minutes and 69 minutes was 28.3%, and 15.2% for durations greater than 70 minutes.

The 46 accidents in the sample set were divided into two parts: 60% of the samples were randomly selected as the training data while the remaining samples were categorized as the testing data. The accidents for model training and model testing were sampled randomly based on the relative frequency of duration.

Table 2. Relative frequency of accident duration

Accident duration (min)	# Samples	Frequency	Cumulative frequency
10~39	26	56.5%	56.5%
40~69	13	28.3%	84.8%
70~109	7	15.2%	100.0%

3.4. Independent variables

An accurate accident duration forecast will assist a driver to decrease uncertainty. The factors influencing an accident are numerous and complex. It is a challenge to accurately predict the impact of an accident due to the uncertainties involved. To capture the phenomenon of accidents, the independent variables incorporated in an accident duration forecasting model were selected from the NPA and NFB databases as shown in Table 3. Most accidents occurred during peak hours (52.2%+21.7%), rainy days (37%), night time (58.7%), at road sections with flexible pavements (95.7%), at road sections with direction facility (e.g., jersey barrier) (65.2%), and as a result of drunk driving (45.7%). Type A1 accidents did not occur during the data collection period. Most variables, such as average upstream speed, average upstream volume, time of day, and weather conditions, can be collected immediately from the database after an accident has been reported to the traffic management centre. This study incorporated all the collected variables to develop the accident duration prediction model and evaluate the model performance. Details are presented in Section 5.

Table 3. Independent variables

Features	Variables	Value	# Samples	%
Average speed at upstream	Average speed at upstream	Continuous variable: km/h		
Average volume at upstream	Average volume at upstream	Continuous variable: # vehicles every 1 min.		
Time of day	Non-peak hours during weekdays	Binary variable: 1: Yes, 0: No	9	19.5
	Peak hours during weekdays	Binary variable: 1: Yes, 0: No	24	52.2
	Non-peak hours during the weekend	Binary variable: 1: Yes, 0: No	3	6.5
	Peak hours during the weekend	Binary variable: 1: Yes, 0: No	10	21.7
Weather condition	Cloudy day	Binary variable: 1: Yes, 0: No	3	6.5
	Rainy day	Binary variable: 1: Yes, 0: No	17	37.0
	Stormy day	Binary variable: 1: Yes, 0: No	1	2.2
Illumination	Day time (excludes the dawn and dusk periods)	Binary variable: 1: Yes, 0: No	18	39.1
	Night time (includes tunnels or underpasses)	Binary variable: 1: Yes, 0: No	27	58.7
Road type (Geographic characteristics)	Tunnel	Binary variable: 1: Yes, 0: No	11	23.9
	Elevated road	Binary variable: 1: Yes, 0: No	4	8.7
# injuries	# injuries	Continuous variable: # injuries passengers		
Accident position	Main lane	Binary variable: 1: Yes, 0: No	5	10.9
	Ramp	Binary variable: 1: Yes, 0: No	4	8.7
	The lane to pass the toll station	Binary variable: 1: Yes, 0: No	1	2.2
Pavement type	Pavement type	Binary variable: 1: Flexible, 0: Rigid	44	95.7
Pavement condition	Pavement condition	Binary variable: 1: Wet, 0: Dry	17	37.0
Obstacle	Obstacle	Binary variable: 1: Yes, 0: No	2	4.3
Direction facility	Direction facility	Binary variable: 1: Yes, 0: No	30	65.2
Collision type	Crash into a roadside parapet	Binary variable: 1: Yes, 0: No	7	15.2
	Overtaking collision	Binary variable: 1: Yes, 0: No	6	13.0
	Crash into a safety island	Binary variable: 1: Yes, 0: No	2	4.3
	Turn over	Binary variable: 1: Yes, 0: No	1	2.2
	Crash into a tree	Binary variable: 1: Yes, 0: No	1	2.2
	Rush out of the road	Binary variable: 1: Yes, 0: No	1	2.2
Causation	Unsafe distance	Binary variable: 1: Yes, 0: No	21	45.7
	Drunk driving	Binary variable: 1: Yes, 0: No	1	2.2
	Changing lanes in an unsafe manner	Binary variable: 1: Yes, 0: No	13	28.3
	Breakdown	Binary variable: 1: Yes, 0: No	3	6.5
	Speeding	Binary variable: 1: Yes, 0: No	2	4.3
	Others	Binary variable: 1: Yes, 0: No	1	2.2
Accident severity	Accident severity (A2: People injured during an accident or died after an accident; A3: Property damage)	Binary variable: 1: A2, 0: A3	5	10.9
Type of involved vehicle	Small truck	Binary variable: 1: Yes, 0: No	12	26.1
	Bus	Binary variable: 1: Yes, 0: No	1	2.2
	Tractor-Semi Trailer	Binary variable: 1: Yes, 0: No	1	2.2
# involved vehicles	# involved vehicles	Continuous variable: # vehicles every 1 min.		

#: the number of

4. Methodology

Based on promising performance in the literature, this study employed two non-parametric machine learning methods, namely the k-nearest neighbour and artificial neural networks, to construct the accident duration prediction model. To prepare relevant data for model development, it was desirable to reduce the dimension of accident features using a correlation analysis for all accidents on National Freeway No. 5.

4.1. Model construction – k-Nearest Neighbour method

The kNN Method is a simple and nonparametric approach for both classification and estimation tasks. This effective method has been widely used in previous studies (Chan et al., 2009; Bustillos et al., 2011; Yu et al., 2011; Chen & Rakha, 2014) for travel time prediction.

The procedure of kNN for estimation is as follows:

(1) The samples are divided into two parts. 60% of samples are used for model training and the remaining 40% of samples are used for model testing. $X_{tr} = \{x_1, x_2, \cdots x_i\}$ and $Y_{tr} = \{y_1, y_2, \cdots y_i\}$ represent the training data sets; x_i denotes the data set of independent variables; and y_i indicates the dependent variables. Meanwhile, $X_{te} = \{x_1, x_2, \cdots x_j\}$ and $Y_{te} = \{y_1, y_2, \cdots y_j\}$ represent the testing data sets; x_j denotes the data set of independent variables; and y_j indicates the dependent variables. Further, i denotes the training data sample and j denotes the testing data sample. Each sample possesses n features.

(2) Calculate the distance between each training data x_i and each testing data x_j using the Euclidean distance shown in Equation 1.

$$d(x_i, x_j) = \sqrt{\sum_{n=1}^{N} (x_i^n - x_j^n)^2} \tag{1}$$

(3) Sort $d(x_i, x_j)$ for each testing data x_j in ascending order and select the first K closest training data set $\{y_1, y_2, \cdots y_k\}$ from $Y_{tr} = \{y_1, y_2, \cdots y_i\}$ for the testing data y_j.

(4) Use the kernel function in Equation 2 to average the first K closest training data set $\{y_1, y_2, \cdots y_k\}$ as the estimated y_j.

$$\hat{y}_j = \frac{\sum_{k=1}^{K} d(x_k, x_j) y_k}{\sum_{k=1}^{K} d(x_k, x_j)} \tag{2}$$

4.2. Model construction – Artificial Neural Networks

The ANN model is a structure describing the complex nonlinear relations between input and output variables. An artificial neuron is a computational model inspired by natural neurons. These consist of inputs, which are multiplied by weights to determine the activation of a neuron. Another function computes the output of the artificial neuron. ANNs combine artificial neurons in order to process information. The ANN model is based on the biological neural system and has been widely applied to prediction and classification problems. The most popular learning algorithm is the back-propagation network.

The ANN algorithm was run with the MATLAB software, which allows continuous and categorical data to be defined clearly. The initial weights from the input layer to the hidden layer were default random values and adjusted in the process until the result was stable and acceptable. In the model structure setting, one hidden layer was chosen and the optimal number of hidden units was determined with the performance index embedded in the software. Before model training, 60% of the samples were randomly selected as the training data and the remaining 40% were selected as the testing data. A typical ANN structure is shown in Figure 2, where the output "y" denotes the accident duration for the accident duration prediction model; the input "x" represents accident and traffic features; "z" stands for the node of the hidden layer; and "w" indicates the weight of the path. The network paradigm is a multilayer perceptron (MLP).

In this study, the number of neurons in the input layer was determined by the features discussed in Section 3.4. The output layer represented the accident duration for the accident duration prediction model.

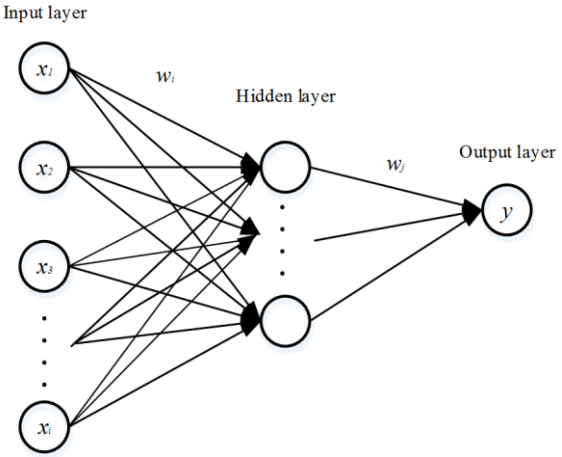


Fig. 2. The structure of an ANN model

4.3. Variable selection – Correlation analysis

To confirm the significant variables and resulting impacts for model construction, all the accident data collected in this study need to be analysed via a correlation analysis. The sample set of accident data collected is a comprehensive collection that covers most of the items recorded in the database. Thus, we first need to specify similar features in order to reduce the complexity of the data set. Relevant literature (Kim & Chang, 2011; Qi & Teng, 2008) indicates that weather conditions, illumination, temporal characteristics, involved vehicle characteristics, and cause of an accident are significant factors in model construction.

Many items for each accident are recorded in the two accident databases. After excluding irrelevant items, correlation analysis is conducted to identify features that have a significant impact on accident duration. The evaluation of accident features is based on the correlation coefficient, *r*. To describe the correlation coefficient, accident duration is defined as the dependent variable *y* and each individual feature is defined as an independent variable *x*.

When an independent variable is both quantitative and continuous, the Pearson correlation coefficient is used to compute the correlation coefficient.

Point-Biserial Correlation is a special case of the Pearson correlation in which the independent variable is a dichotomous variable. When the independent variable is a binary variable, the correlation coefficient is computed by Equation 3 as follows:

$$r_{xy} = r_{pb} = \frac{M_1 - M_0}{S_y} \times \sqrt{\frac{n_1 n_0}{n(n-1)}} = \frac{M_1 - M_0}{\sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}} \times \sqrt{\frac{n_1 n_0}{n(n-1)}} \tag{3}$$

where:

- r denotes the correlation coefficient;
- M_1 represents the mean value of y when the value of the independent variable x is 1;
- M_0 indicates the mean value of y when the value of the independent variable x is 0;
- n_1 denotes the number of independent variables x whose values are 1;
- n_0 represents the number of independent variables x whose values are 0.

The results of the correlation analysis for the accident duration prediction model are presented in Table 4. The correlation analysis yielded a high significance at a confidence level of 99% for the following variables: Average volume at upstream, the number of injuries, crash into roadside parapets, and accident severity. An additional five variables, namely rainy day, day time, pavement condition, drunk driving, and the number of vehicles involved, significantly correlated with accident duration at a confidence level of 95%.

Table 4. Correlation coefficients of accident duration and features

Features	Variables	Coeff.	p-values
Average speed at upstream	Average speed at upstream	-0.107	0.478
Average volume at upstream	Average volume at upstream	-0.479 **	0.001
Time of day	Peak hours during weekdays	-0.142	0.347
	Non-peak hours during the weekend	0.156	0.301
	Peak hours during the weekend	-0.155	0.303
Weather condition	Cloudy day	0.179	0.235
	Rainy day	0.305 *	0.039
	Stormy day	0.063	0.677
Illumination	Day time (excludes dawn and dusk periods)	-0.291 *	0.049
	Night time (includes tunnels or underpasses)	0.278	0.062
Road type	Tunnel	-0.069	0.647
(Geographic characteristics)	Elevated road	0.374	0.100
# injuries	# injuries	0.409 **	0.005
Accident position	Main lane	-0.195	0.195
	Ramp	-0.214	0.153
	The lane to pass a toll station	0.152	0.313
Pavement type	Pavement type	0.002	0.987
Pavement condition	Pavement condition	0.314 *	0.034
Obstacle	Obstacle	0.052	0.732
Direction facility	Direction facility	0.130	0.387
Collision type	Crash into a roadside parapet	0.457 **	0.001
	Overtaking collision	0.066	0.665
	Crash into a safety island	0.058	0.702
	Turn over	0.040	0.791
	Crash into a tree	0.083	0.584
	Rush out of the road	0.152	0.313
Causation	Unsafe distance	0.345	0.109
	Drunk driving	0.329 *	0.026
	Changing lanes in an unsafe manner	0.124	0.411
	Breakdown	0.090	0.551
	Speeding	0.075	0.620
	Others	0.219	0.143
Accident severity	Accident severity (A2: People injured during an accident or died after an accident; A3: Property damage)	0.518 **	0.000
Type of involved vehicle	Small truck	0.172	0.253
	Bus	0.032	0.831
	Tractor-Semi Trailer	0.160	0.289
# involved vehicles	# involved vehicles	-0.282 *	0.038

#: the number of
*: indicates that the independent variable significantly correlated with accident duration at a confidence level of 95%.
**: indicates that the independent variable significantly correlated with accident duration at a confidence level of 99%.

4.4. Performance evaluation

Evaluation of model accuracy is required in order to assess the performance of the prediction models. The mean absolute percentage error (MAPE) is a summary measure widely used for evaluating the accuracy of prediction results (Zhan et al., 2011; Khattak et al., 2012; Li, 2015; Dimitriou & Vlahogianni, 2015; Chung, 2010; Li et al., 2015; Li et al., 2016). MAPE was applied in this study to fairly compare relative performance among various model settings.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{V_i^{actual} - V_i^{predicted}}{V_i^{actual}}\right| \times 100\% \quad (4)$$

where:

- V_i^{actual} denotes the actual value of observation;
- $V_i^{predicted}$ represents the predicted value of observation;
- n indicates the sample size.

The lower the MAPE value is, the more accurate the prediction model will be. A MAPE value less than 10% indicates a highly accurate prediction; a MAPE value between 11% and 20% means a good prediction; and a MAPE value between 21% and 50% refers to a reasonable prediction. The threshold of MAPE was suggested by Lewis (1982).

5. Evaluation

Since the fundamental theory of the kNN and ANN training algorithms are stochastic-oriented, various combinations of initial weights and hidden units may lead to different states of convergence. The literature, however, does not offer a general guideline as to determining the best choice. Therefore, a suitable number of trials should be implemented to verify the performance of the proposed kNN and ANN models.

Given the sample set of 46 accidents prescreened with the corresponding accident duration times, ten experiments were conducted for examining the proposed methodology. In each experiment, 60% of the data was randomly selected as the training set from the sample set and the remaining 40% of data served as the testing set.

Four accident duration prediction models were developed in this study. Both Models 1 and 2 utilized the kNN method as the key algorithm. Model 1 incorporated all the variables in Table 3 while Model 2 incorporated the significant variables in Table 3. Both Models 3 and 4 used the ANN method as the key algorithm. Model 3 incorporated all the variables in Table 3 and Model 4 incorporated the significant variables in Table 3. The same training/testing set was applied to the four models in each experiment.

5.1. Results of the accident duration prediction model

Table 5 depicts the results of the ten experiments. The average MAPE values were below 48% for each type of model, yielding a level of reasonable prediction. For most experiments, Model 4 which applied the ANN method and incorporated the significant variables provided the best prediction results and the MAPE values close to 20%. Based on the results of model evaluation, Model 4 can provide good and reasonable predictions. The performances of the model that incorporated significant variables were better than those that incorporated all the variables. The models that applied the ANN method could predict the accident duration more accurately than those that applied the kNN method.

Table 5. The MAPE values of the accident duration prediction model

	Model 1	**Model 2**	**Model 3**	**Model 4**
Method for model construction	kNN	kNN	ANN	ANN
Independent variables in model	All variables	Significant variables	All variables	Significant variables
Experiments				
1	29.3%	28.8%	25.6%	20.9%
2	41.1%	40.4%	31.7%	21.7%
3	63.8%	57.6%	28.2%	20.1%
4	53.6%	45.3%	30.6%	21.5%
5	49.4%	44.9%	31.6%	21.9%
6	48.8%	46.1%	30.7%	23.7%
7	52.4%	49.8%	31.8%	26.7%
8	34.4%	40.6%	33.4%	29.9%
9	50.5%	57.0%	33.8%	33.8%
10	49.3%	45.9%	31.5%	34.2%
Average	47.3%	45.6%	30.9%	25.4%

Table 6 lists the performance difference among the four models for a further comparison of model performance. The first column of Table 6 compares the performance of the two models which applied the kNN method. The average performance of Model 2 was slightly better than the average performance of Model 1 and the average improvement was about 2.4%. The second column of Table 6 compares the performance of the two models which applied the ANN method. The average performance of Model 4 was better than the average performance of Model 3 and the improvement was about 18.4%. This result indicates that Model 2 and Model 4 may be considered a potential candidate approach to predict accident duration when a suitable set of accident variables is provided.

The third column of Table 6 compares the performance of the two models which incorporated all the variables. The average performance of Model

3 was better than the average performance of Model 1 and the average improvement was about 34.7%. The fourth column of Table 6 compares the performance of the two models which incorporated significant variables. Similarly, the average performance of Model 4 was better than the average performance of Model 2 and the average improvement was about 44.3%. Based on the aforementioned results, the ANN method is more efficient in developing the relationship between traffic/accident data and accident duration than the kNN method when the models incorporate the same variables.

Figure 3 shows the performance assessment with respect to the predicted accident duration vs. actual accident duration. Generally speaking, most data points are scattered along the 45° line with a reasonable distance (discrepancy), especially the plots of Model 4.

Table 6. The difference of model performance

Experiments	Performance improvement from Model 1 to Model 2	Performance improvement from Model 3 to Model 4	Performance improvement from Model 1 to Model 3	Performance improvement from Model 2 to Model 4
1	1.6%	18.3%	12.8%	27.6%
2	1.6%	31.5%	22.9%	46.4%
3	9.8%	28.5%	55.9%	65.0%
4	15.6%	29.8%	43.0%	52.5%
5	9.2%	30.7%	36.1%	51.3%
6	5.5%	22.9%	37.0%	48.6%
7	5.0%	15.9%	39.3%	46.3%
8	-18.1%	10.5%	2.8%	26.4%
9	-12.8%	0.2%	33.0%	40.7%
10	6.9%	-8.8%	36.2%	25.4%
Average	2.4%	18.0%	34.7%	44.3%

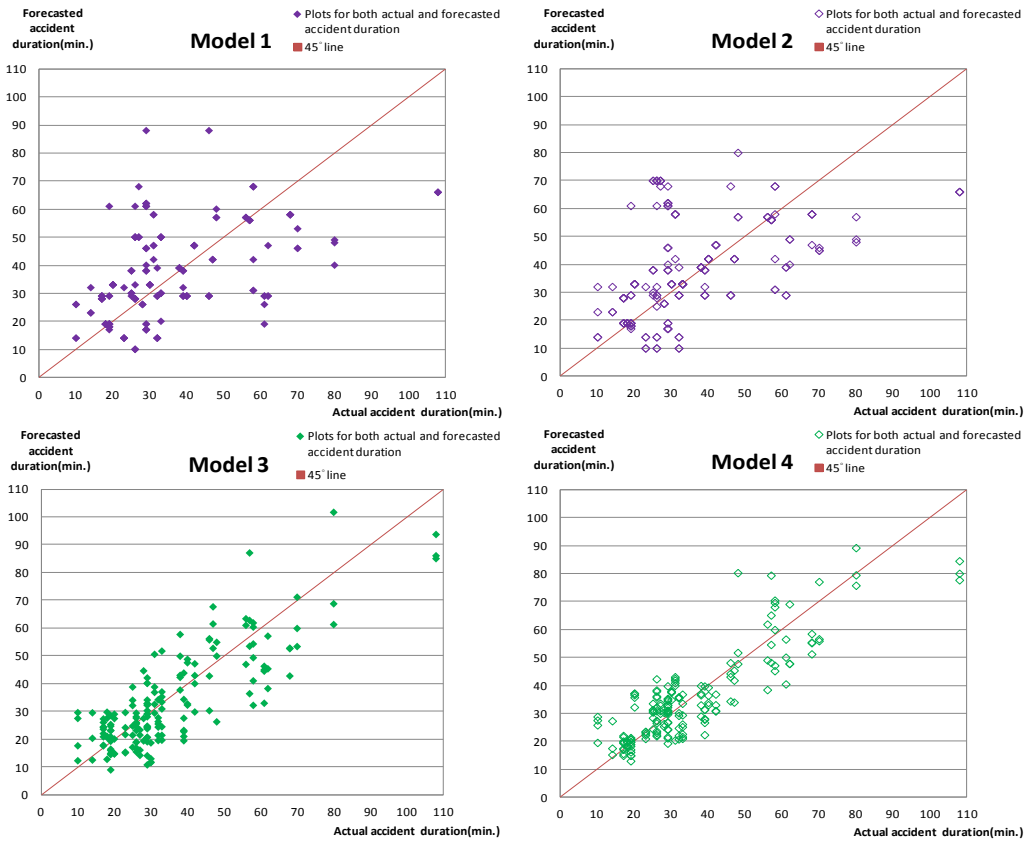


Fig. 3. Assessment results of the four models

As can be seen, most plots of Model 2 are closer to the 45° line than most plots of Model 1; meanwhile, most plots of Model 4 are closer to the 45° line than most plots of Model 3. This indicates that the prediction of models that incorporate significant variables can match the actual accident duration.
Most plots of Model 3 are closer to the 45° line than most plots of Model 1; meanwhile, most plots of Model 4 are closer to the 45° line than most plots of Model 2. This implies that the models that apply the ANN method may sufficiently capture the relationship between the inputs (accident features) and the output (accident duration).

5.2. Performance comparison for circumstances

Table 7 shows the MAPE values and p-values of a t-test for two circumstances of each feature for a comparison of the prediction performance. The statistic t-test was used to test the equality of MAPE values for two circumstances of each feature.
The p-values were greater than 0.05 and the MAPE values for the four models were not significantly different between (i) rainy days and other weather conditions, (ii) dry and wet pavement conditions, and (iii) a crash into a roadside parapet and other collision type. This result means that the four models can provide a similar prediction performance in all types of weather conditions, pavement conditions, and collision type.
Based on the p-values below 0.05, Model 4, which incorporated significant variables and employed the ANN method, can provide a more accurate prediction of accident duration when the circumstances involved the day time or drunk driving than those that involved night time and did not involve drunk driving.
For Model 1 and Model 2, the p-values between Type A2 and Type A3 accidents were less than 0.05 and the MAPE values of the Type A2 accident were significantly lower than those of the Type A3 accident. This result shows that the models that apply the kNN method may better capture the phenomenon of the Type A2 accident than that of Type A3.

Table 7. A comparison of model performance by circumstance

			Model 1	**Model 2**	**Model 3**	**Model 4**
		Method for model construction	**kNN**	**kNN**	**ANN**	**ANN**
Features	**Circumstances**	**Independent variables in model**	**All variables**	**Significant variables**	**All variables**	**Significant variables**
Weather	Rainy day	MAPE	39.3%	46.2%	27.3%	30.3%
	Others	MAPE	51.2%	45.4%	32.6%	23.1%
	t-test for the above two circumstances	p-value	0.118	0.920	0.242	0.109
Illumination	Day time	MAPE	54.9%	45.0%	29.2%	**17.8%**
	Night time	MAPE	42.1%	46.1%	32.1%	30.6%
	t-test for the above two circumstances	p-value	0.077	0.894	0.501	0.003
Pavement condition	Dry	MAPE	49.2%	46.0%	32.6%	23.3%
	Wet	MAPE	43.6%	45.0%	27.5%	30.0%
	t-test for the above two circumstances	p-value	0.461	0.903	0.249	0.160
Collision type	Crash into roadside parapet	MAPE	49.9%	52.3%	28.0%	24.6%
	Others	MAPE	46.8%	44.5%	31.4%	25.6%
	t-test for the above two circumstances	p-value	0.763	0.470	0.571	0.864
Drunk driving	Yes	MAPE	51.3%	44.2%	38.2%	**5.8%**
	No	MAPE	47.2%	45.7%	30.8%	25.8%
	t-test for the above two circumstances	p-value	0.884	0.960	0.652	0.226
Accident severity	A2	MAPE	**21.0%**	**19.8%**	31.7%	22.4%
	A3	MAPE	50.2%	48.5%	30.8%	25.8%
	t-test for the above two circumstances	p-value	0.013	0.022	0.895	0.632

6. Conclusions

An accident data set with 46 cases recorded in two databases during 2012 was built for the accident duration prediction models. The k-nearest neighbour (kNN) and artificial neural network (ANN) approaches were then employed to develop the prediction model when the relevant information regarding accident features/variables were provided. Before model development, a correlation analysis was applied to reduce the scale of interrelated features/variables. Based on the correlation analysis results, Average volume at upstream, the number of injuries, crash into roadside parapets, accident severity, rainy day, day time, pavement condition, drunk driving, and the number of vehicles involved significantly correlated with accident duration. The primary features identified on the case site were consistent with those reported in the literature. The evaluation results of prediction models indicate that the proposed ANN approach is promising as numerical experiments yielded good and reasonable performance in various model compositions based on mean absolute percent error values. The prediction performance can be improved, when the prediction model selected the significant variables and reduced variables dimension. Accurately forecasted accident duration will assist a driver to decrease uncertainty.

For future studies, more accident data should be collected and processed to facilitate the learning capability of the proposed models. Since default settings were mostly used for the time being, a number of parameters needed to be carefully set to further enhance the training mechanism.

References

[1] BUSTILLOS, B. I., CHIU, Y. C., 2011. Real-time freeway-experienced travel time prediction using N-curve and K nearest neighbor methods, *Transportation Research Record*, 2243, pp. 127–137.

[2] CHAN, K. S., LAM, W. H. K., TAM, M. L., 2009. Real-time estimation of arterial travel times with spatial travel time covariance relationships, *Transportation Research Record*, 2121, pp. 102–109.

[3] CHEN, H., RAKHA, H. A., 2014. Real-time travel time prediction using particle filtering with a non-explicit state-transition model, *Transportation Research Part C*, 43 (1), pp. 112–126.

[4] CHIEN, I. J., DING, Y., WEI, C. H., 2002. Dynamic Bus Arrival Time Prediction with Artificial Neural Networks, *Journal of Transportation Engineering*, 128(5), pp. 429–438.CHOI, H. K., 1996. Predicting Freeway Traffic Incident Duration an Expert System Context Using Fuzzy Logic, PhD Dissertation, University of Southern California, Los Angeles, USA.

[5] CHUNG, Y. S., CHIOU, Y. C., LIN, C. H., 2015. Simultaneous equation modeling of freeway accident duration and lanes blocked, *Analytic Methods in Accident Research*, 7, pp. 16–28.

[6] CHUNG, Y., 2010. Development of an accident duration prediction model on the Korean Freeway Systems', *Accident Analysis Preview*, 42(1), pp. 282–289.

[7] DIMITRIOU, L., VLAHOGIANNI, E. I., 2015. Fuzzy modeling of freeway accident duration with rainfall and traffic flow interactions, *Analytic Methods in Accident Research*, 5-6, pp. 59–71.

[8] GARIB, A., RADWAN, A.E., AL-DEEK, H., 1997. Estimating magnitude and duration of incident delays, *Journal of Transportation Engineering*, 123(6), pp. 459–466.

[9] GUO, B., NIXON, M. S., 2009. Gait feature subset selection by mutual information, *IEEE Transactions on Systems, Man, and Cybernetics- Part A: Systems and Humans*, 39 (1), pp. 36–46.

[10] HOJATI, A. T., FERREIRA, L., WASHINGTON, S., CHARLES, P., 2013. Hazard based models for freeway traffic incident duration, *Accident Analysis and Prevention*, 52, pp. 171–181.

[11] KHATTAK, A., WANG, X., ZHANG H., 2012. Incident management integration tool: dynamically predicting incident durations, secondary incident occurrence and incident delays, *IET Intelligent Transport Systems*, 6(2), pp. 204–314.

[12] KIM, W., CHANG, G. L., 2011. Development of a Hybrid Prediction Model for Freeway Incident Duration: A Case Study in Maryland. *International Journal of Intelligent Transportation Systems Research*, 10(1), pp. 22–33.

[13] LEE, Y., WEI, C. H., 2009. Freeway travel time forecast using Artificial Neural Networks with Cluster Method, *12th International Conference on Information Fusion*, pp. 1331–1338.
[14] LEWIS, C. D., 1982. *Industrial and Business Forecasting Method*. London: Butterworth Scientific.
[15] LI, R., 2015. Traffic incident duration analysis and prediction models based on survival analysis approach, *IET Intelligent Transport Systems*, 9(4), pp. 351–358.
[16] LI, R., PEREIRA, F. C., BEN-AKIVA, M. E., 2015. Competing risks mixture model for traffic incident duration prediction, *Accident analysis and prevention*, 75, pp, 192–201.
[17] LI, T., YANG, Y., WANG, Y., CHEN, C., YAO, J., 2016. Traffic fatalities prediction based on support vector machine, *Archives of Transport*, 39(3), pp. 21–30.
[18] NAM, D., MANNERING F., 2000. An exploratory hazard-based analysis of highway incident duration, *Transportation Research Part A*, 34(2), pp. 85–102.
[19] Pamula, T., 2012. Classification and Prediction of Traffic Flow Based on Real Data Using Neural Networks, *Archives of Transport*, 24(4), pp. 519–529.
[20] QI, Y., TENG, H., 2008. An Information-Based Time Sequential Approach to Online Incident Duration Prediction, *Journal of Intelligent Transportation Systems*, 12(1), pp. 1–12.
[21] SMITH, K., SMITH, B. 2001. Forecasting the Clearance Time of Freeway Accidents, Research Report STL-2001-01. Center for Transportation Studies, University of Virginia, Charlottesville, VA.
[22] Spławińska, M., 2015. Development of models for determining the traffic volume for the analysis of roads efficiency, *Archives of Transport*, 33(1), pp. 81–91.
[23] VALENTI, G., LELLI, M., CUCINA, D., 2010. A comparative study of models for the incident duration prediction, *European Transport Research Review*, 2(2), pp. 103–111.
[24] VLAHOGIANNI, E. I., KARLAFTIS, M. G., 2013. Fuzzy-entropy neural network freeway incident duration modeling with single and competing uncertainties, *Computer-Aided Civil and Infrastructure Engineering*, 28(6), pp. 420–433.
[25] WANG, W., CHEN, H., BELL, M. C., 2005. Vehicle breakdown duration modeling', *Journal of Transportation and Statistics*, 8(1), pp. 75–84.
[26] WINSTON, C., LANGER, A., 2006. The effect of government highway spending on road users' congestion costs. *Journal of Urban Economics*, 60(3), pp. 463–483.
[27] YU, B., LAM, W. H. K., TAM, M. L., 2011. Bus arrival time prediction at bus stop with multiple routes', *Transportation Research Part C*, 19(6), pp. 1157–1170.
[28] ZHAN, C., GAN, A., HADI, M., 2011. Prediction of lane clearance time of freeway incidents using the M5P tree algorithm. *IEEE Transactions on Intelligent Transportation Systems*, 12(4), pp. 1549–1557.
[29] ZHANG, H. M., 2000. Recursive prediction of traffic conditions with neural network models', *Journal of Transportation Engineering*, 126(6), pp. 472–481.

www.czasopisma.pan.pl PAN POLSKA AKADEMIA NAUK www.journals.pan.pl
ARCHIVES OF TRANSPORT ***ISSN (print):*** *0866-9546*
Volume 43, Issue 3, 2017 ***e-ISSN (online):*** *2300-8830*
DOI: *10.5604/01.3001.0010.4229*

TOPOLOGY-BASED APPROACH TO THE MODERNIZATION OF TRANSPORT AND LOGISTICS SYSTEMS WITH HYBRID ARCHITECTURE. PART 1. PROOF-OF-CONCEPT STUDY

Iouri N. Semenov[1], Ludmiła Filina-Dawidowicz[2]

[1,2] West Pomeranian University of Technology Szczecin, Faculty of Maritime Technology and Transport, Szczecin, Poland

[1]e-mail: jusiem@zut.edu.pl
[2]e-mail: lufilina@zut.edu.pl

Abstract: *Industrial companies are linked with affiliated firms, suppliers and customers through the supply chains. Such chains operate within large-scale networks directly related to distribution and warehousing. In order to meet the market demands and customer new expectations, the various components of the supply chains have to be developed i.a. through implementation of the innovative vehicles, green and blue technologies. Moreover, modernization processes of transport and logistics system need to be resistant to crucial mistakes related to innovative solutions implementation in order to exclude domino effect occurrence. The authors attempt to build topology-based approach to the modernization of transport and logistics systems. It is assumed that each innovation application is the independent element-based coalition, possessing linked object structure. The results of multi-year researches demonstrate the offered approach as a useful tool to analyze innovative changes for obsolete transport and logistics system as hybrid structure. The ways of system structure transformation, as well as possible innovative effects are considered. The preliminary results have been obtained for compositions on meso-level for the case of marine propulsion modernisation. The paper is illustrated by various examples.*

Key words: *topology-based approach, transport and logistics system, innovative changes, component-based coalition.*

1. Introduction

As a result of progressive evolution, mobility remains the key to prerequisite for prosperity of the society. Long before the industrial revolution, our civilisation had evolved due to innovative changes. Such alterations had developed from a propulsive force of wind and sail, animal power and wheel up to a coal-fired propulsive plant and steam engine, as well as later on petroleum and Diesel ones.

Therefore, transport systems transformations have taken place within more than 4000 years, beginning from scattered, uncoordinated and homogeneous (unimodal) structures, e.g. mail-coaches or sailing fleet, up to networked, coordinated and heterogeneous (hybrid) structures, hereinafter referred to TLS (*Transport and Logistics System*).

The authors examined innovations covering the period from 1715 to 2015 that were divided into three categories (Fig. 1). These categories may be named as:

1. Incremental Innovations (improvement). These innovations consist of small, yet meaningful improvements.
2. Substantial Innovations (breakthrough). There are meaningful changes that give consumers some demonstrably new features.
3. Transformational Innovations (disruptive). This kind of innovations often eliminates existing transport means or totally transforms them. Conducted researches show that transformational innovations tend to be championed by those who aren't wedded to obsolete structures (clipper, chariot, steam locomotive etc.).

After industrial revolution obsolete vehicles were eliminated and replaced by engine-based means. In XX-century the idea to integrate such vehicles within intermodal transport systems, defined as hybrid transport systems, was created (Konings et al., 2005) (Fig. 2).

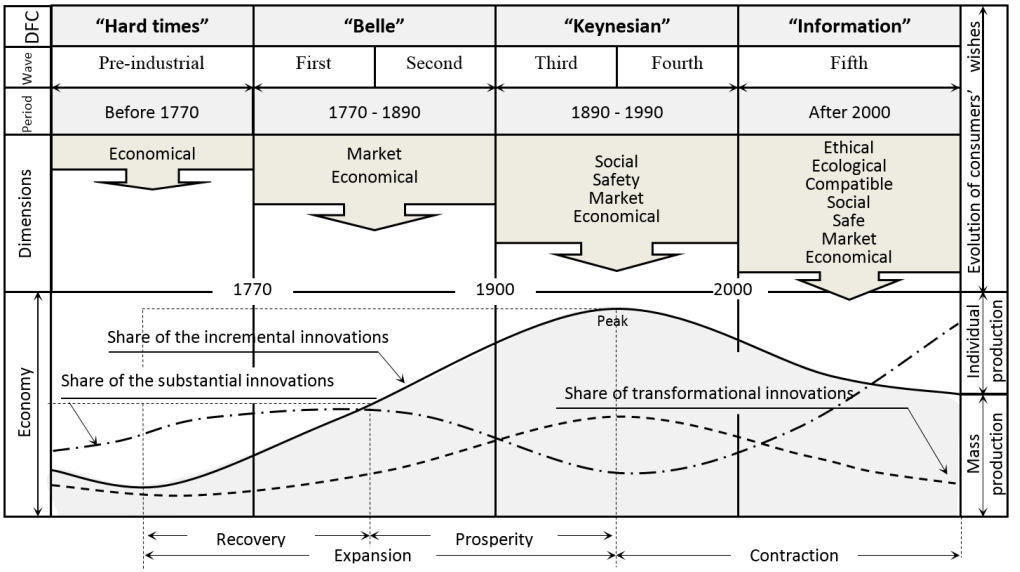


Fig. 1. Worldwide changes at Transport/Logistics innovations
Source: Authors' research on the basis of Aho et al. (2006); Bolt and van Zanden (2014); United Nations (2015)

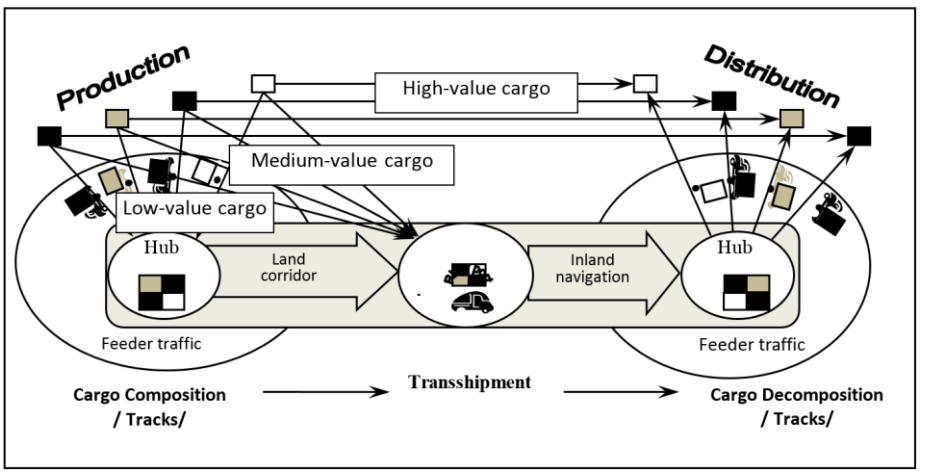


Fig. 2. An example of the intermodal Transport and Logistics System

Intermodal transport links different participants (e.g. shipper, carrier, consignee) through supply chain development (Jacyna-Gołda, 2015; Kerbache and Maccregor Smith, 2004). Conducted literature analysis revealed that TLS problems were widely described in current literature (Chataway et al., 2014; Hölzl and Janger, 2011).

The drivers and sources of innovations have been examined e.g. by (Dosi et al., 2000; Chaminade et al., 2010, Schumpeter, 1950) paying particular attention to the technological and organizational companies resources. Moreover, it is known that implementation of innovation does not always have positive effects. Sometimes the rejection or the lack

of compatibility with the existing system takes place. This phenomenon generates fundamental transport or logistics problem, i.e. the need of combinatorial optimization of the TLS assembled by CSG (C*oalition Structure Generation*) techniques (Mauro et al., 2010; Rahwan et al., 2009).
Improvement of any TLS development may be based on the multi-agent approach (Chen and Cheng, 2010; Chen and Wang, 2009; Davidsson et al., 2005; Graudina and Grundspenkis, 2005; Lin, 2011, Modelewski and Siergiejczyk, 2013; Rocha et al., 2014) and object-oriented approach (Arm Badr-El-Din, 2013; Crespi et al., 2008; Juman et al., 2013).
There is a number of publications examining the use of coalition structure in systems transformations (Aziz and de Keijzer, 2011; Baras, 2011; Semenov, 2006; Voice et al., 2012). Nevertheless, despite extensive research in this field there are still gaps referring to the lack of a uniform approach to TLS modernization based on the interaction of innovative and obsolete coalitions that can be considered as independent object modules.
The article aims to demonstrate the topological approach to analyze innovative changes for obsolete TLS as hybrid system.

2. The levels of modernization process

The TLS modernization has never been the simple and easy process. According to O. Levander (Rolls-Royce): "*It has never been more difficult to know what's the right investment decision to make when selecting a new vessel. The answer is flexibility – the 'future-proof' ship*" (Low-cost, 2017). Nowadays, the TLS development stage is leading to the substitution of obsolete and polluting technologies by modern and ecological ones, as well as mobility improvement through its inner openness to innovative changes. Therefore, the design of such systems should be based on risk management of each component, coalition and a whole composition, as well as inclusion of modernisation process in the context of maximising efficiency and minimising disturbances.
For that reason, all TLSs can be divided into two groups:
A Group: **so-called COTLSs *(Closed TLSs).*** Each system is a closed structure, if it isn't connected with defined environment (Hubka and Eder, 1988). According to the system approach, the closed system is isolated form of compositions, which must be developed as high-reliable construction. Such systems should have higher assurance factor and survival rate. They stand out by scarcity of evolutionary mechanisms, and consequently are ill-adapted to any transformations. The striking examples are the subsea oil and gas pipelines, as well as the TLPs (Tension Leg Platforms) transformed from the jack-up drilling rigs.
B Group: so-called OTLSs *(Opened TLSs)*. Each system is an opened multi-element structure, if it is connected with environment by at least one input or/and output. The feasibility of the OTLS transformations depends on a degree of such system worsens, structural complexity and modernisation tasks.
The OTLS improvement can be carried out at micro-, meso- and mega-level of TLS hierarchy. Let's consider these levels closely.

1) The micro-level: describes transport systems at low aggregation levels and refers to the functioning elements of systems and is, therefore, a valuable assessment instrument for innovative analysis, see e.g. Semenov (2008). An engine is the classic example of the component-based coalition embedded into any kind of engine-based vehicle. Its improvement implies modification of one or several components, including the camshaft, the crank shaft, the flywheel, cylinders, pistons, etc., as well as a principle of their interactions. An assembly of such components can be formed in different ways. As a result various MAGI (*Micro-Areas of Geometrical Incompatibility*) are generated (Bhalla et al., 2014). Any innovative micro-level transformations have the tendency to complicate the engineering systems drastically, increasing the risk of costly human errors occurrence. This fact stimulates the reduction of total number of obsolete coalitions within such compositions, raising its reliability and usability. One of the representative examples is the innovative solution developed by Bosch Corp. uniting the few components within coalition "*Combined Active and Passive Safety*" (Bosch, 2017).

2) The meso–level is wedged between the macro- and the micro-levels. Therefore, the meso-level describes the transport system from an intermediate aggregation level, and this type of analysis acknowledges the mutual coherence of actors' groups (Schenk et al., 2007). A vehicle is a good example of the next OTLS ordered level.

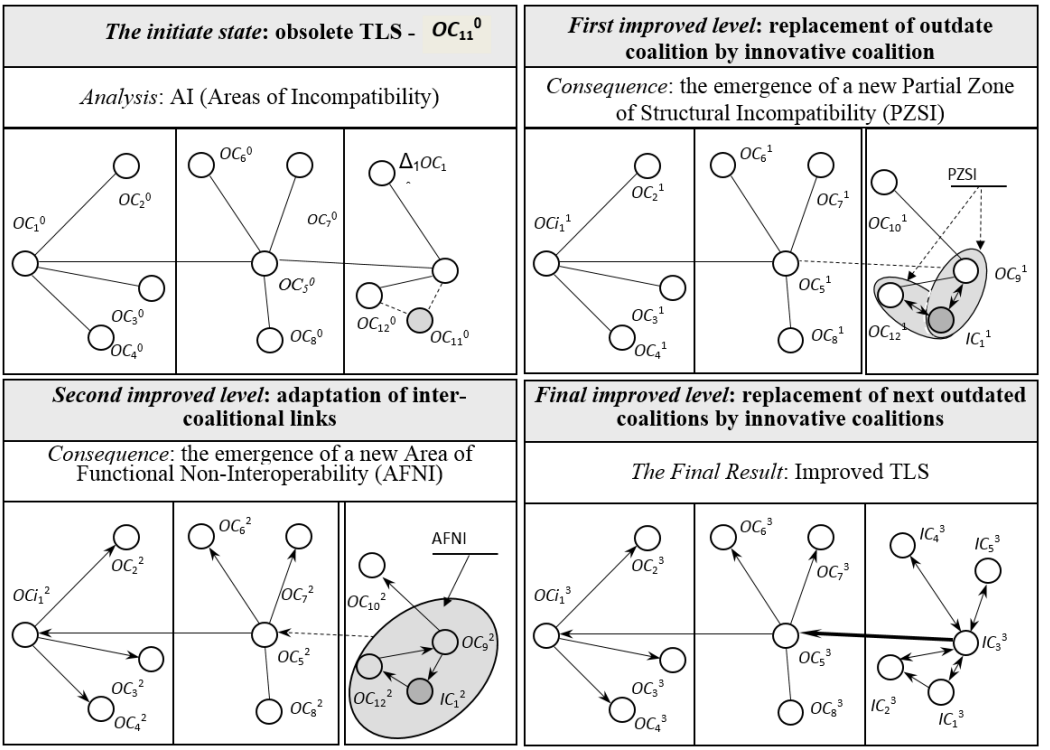


Fig. 3. The stages of outdated TLS modernization, where OC – outdated coalition, IC – innovative coalition

Improvement of the vehicle implies modification of various coalitions (subsystems), i.a.:

- safety subsystems including anti-lock braking subsystem,
- subsystems preventing roll-overs and skids,
- a ship hull, an power-plant, an engine, a propulsion devices etc.

Consequently, assemblies of aforesaid coalitions can be formed in different ways (Fig.3) and as a result, MAGI could generate PZSI (*Partial Zones of Structural Incompatibility*). Moreover, the upgrade within single coalition, as a rule, leads to corresponding alterations in other coalitions. For example, invention of a hybrid engine demanded a reconfiguration of car bodies, and usage of transverse under floor installation of car engine. In addition, consumers are heterogeneous in their needs, opportunities, and wishes. In conclusion, each transport system is struggling with various AFNI (*Areas of Functional Non-Interoperability*).

To escape from these difficulties, the managers developed and implemented novelties under imperative motto "the preferences declared by the majority of customers should be above all". As a rule, it is converged with aspiration of persistent reduction of the vehicles prices and services expenses. Unfortunately, often enough the satisfaction of such wishes can be achieved only by a refrain from innovative transformations and, as a result, application of outdated solutions, e.g. usage of petrol engines in car production. Also the contrary situation may take place, when the clients' wishes concern an environmental protection, reliability, safety or comfort. Managers should take into account that the AI (*Areas of Incompatibility*), both functional and structural, can arise in each case.

The striking example of described case concerns unsuccessful decisions taken during the design of several VLCCs (*Very Large Crude Carriers*) decks. The critical imitation of helicopters landing space and the limited opportunity of emergency crew

escape took place. However, history shows that formation of AI does not finish the progress. The development of the sailing fleet contains enlightening examples of cut-and-try method.

Furthermore, rapid growth of the European economy and reinforcement of trade relations in both South and North America within XVI-XVII centuries resulted in the need for drastic reduction of the transportation time. This problem had been solved owing to transformations of sailing equipment, including permanent changes of ship's hull forms, localisations and a quantity of sails and, consequently, number of sail masts.

These transformations caused the increase in the ship's centre of KG (*Gravity above Keel*), and loss of static and dynamic stability of sailing ships, causing negative effects, i.a:

- sailing ships had to take aboard of huge quantity of the solid ballast,
- necessity to increase a crew size and as a result to expand the amount of supplies (fresh water and foods) on ships that reduced cargo capacity.

However, growing demand for freight caused upsizing of sailing ships, both the number of sail masts and crews increased aiming to support the transportation capacity. Similar changes took place until the end of the XIX century, giving the chance for steamships fleet development (Fig. 4).

Today shipping industry has become a key component of the world's economy. Over 90% of global trade is carried by sea. The world fleet of sea-going merchant ships reaches over 104,000 ships. As a result, such problems as ship CO_2 emissions and sulphur emissions occur (Table 1). From 2015, ships operating in SECA (*Sulphur Emissions Control Areas*) have to use fuels with 0.1% or less sulphur content. This sulphur regulation put pressure on ship owners and operators forcing them to invest in cleaner fuels and green technology, as well as ships innovations. For that reason, during the last decade an increasing focus on emissions reduction for new and existing ships had been observed. Therefore, ship-owners optimise the form of ship hulls, fit equipment for emission reduction, install new propellers and tune engines.

Table 1. Outcome-oriented goals for modern fleet of merchant ships

Outcome-oriented goals	**Description**
Market development	Improve the competitive advantage for shipping companies through implementation of green technologies
Health protection	Reduce premature deaths from exposure to particulate emissions *
Climate preservation	Reduce per-capita CO_2 emissions from ferries and merchant ships
Technological improvement	Improve shipping sustainability through usage of good practices and innovative solutions
Law compliance	Meeting the SECA requirements**
Profitability ensuring	The ability to finance ongoing operations and future fleet growth
* WHO reports: in 2012 around 7 mln people died as a result of air pollution (one in eight of total global deaths). ** Sulphur Oxides, International Maritime Organization, 2014, retrieved 4 May 2014.	

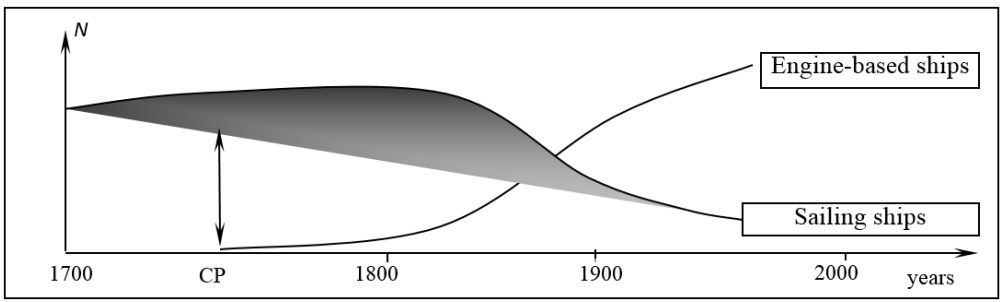


Fig. 4. The global fleet development: Engine-Based vs. Sailing Fleet
N-the number of worldwide ships, CP-Critical Point in civilization development (the "Belle" Epoch - Fig. 1)

This situation influences on production of emission-efficient ships using new solutions, such as exhaust gas cleaning (scrubbers) systems or ships adapted for the LNG fuel. Coming new solutions include the combination of three different technologies: water injection, SCR (*Selective Catalytic Reduction*), as well as EGR (*Exhaust Gas Recirculation*). The EGR technology is fairly straightforward in on-land applications, while recycling exhaust gases from marine fuel back into an engine causes different challenges.

3) The macro level relies on a very aggregate view of transport systems and determines modernisation processes at meso- and micro-levels and considers systems as hybrid structures. Therefore, each harbour is the representative example of the OTLS at the macro level because consists of the multi-coalition composition, which includes (Fig. 5):

- unmovable components, including road/railway networks, storage warehouses,
- mobile components, i.a. the portal cranes, wheel loaders.

Seaports modernisation faces the number of possible complications, dealing with regulations changes, technological incompatibility with simultaneous widening of services spectrum etc. Therefore, each unfortunate modernisation can cause significant expenditure and even investment fiasco.

3. The modernization process

3.1. Core assumptions and tasks

Existing TLS are characterised both by a wide variety of structures (from opened to fully closed complexes), and different purposes. Therefore, we'll introduce a few conditions for modernization and development of OTLS. The major conditions are as follows:

- a phased transformation process is based on unified procedures;
- an irreversible transformation process that once started modernization cannot be stopped;
- a progressive transformation process is not only inevitable but desirable;
- a risky transformation process creates on-target, as well as off-target effects;
- a transformation process changes the obsolete system into a modernity state;
- a pushed transformation process builds chance for obsolete system (Fig. 6).

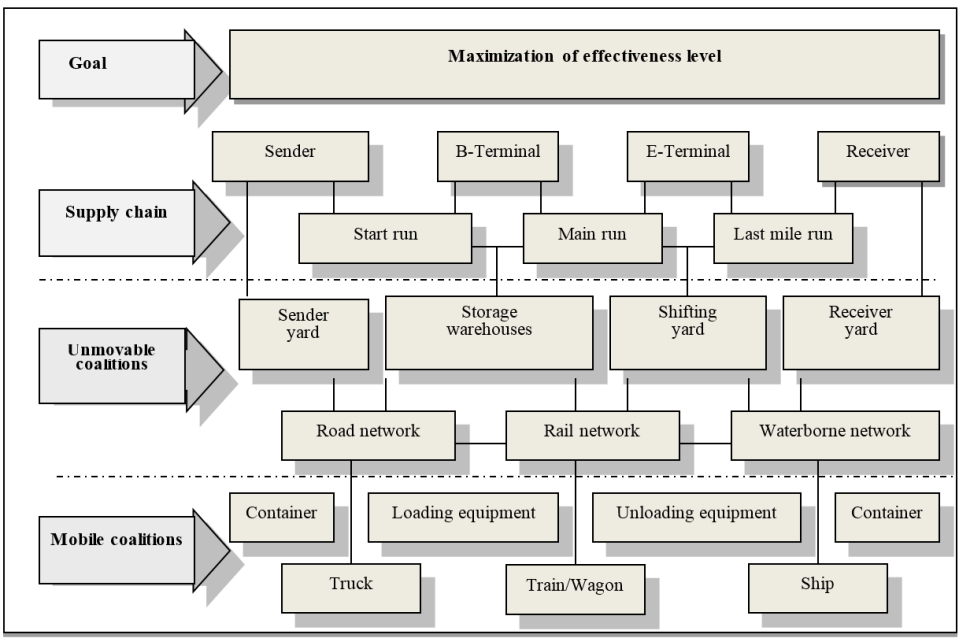


Fig. 5. The seaport as the Hybrid OTLS

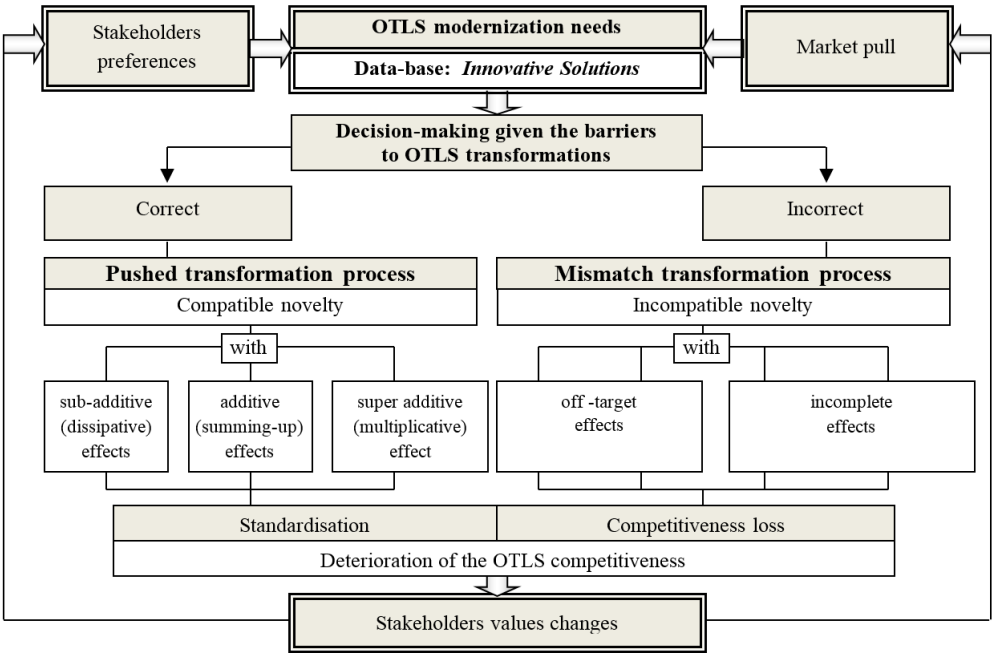


Fig. 6. Conceptual innovative sequence of the OTLS transformation

The presented approach should help to reduce modernization risk or destruction threats for renovated system as a result of the imprudent innovative decisions.

Consequently, the process of modernising each TLS begins with the development of a concept, core assumptions and general goals that forms innovative investment decisions. Setting of clearly defined goals will help to create the basic procedures.

3.2. The modernization principles

The OTLS modernization processes are based on two rules. First, each OTLS should be assembled by CSG techniques and connected together in order to carry out the set functions. Second, the achieved results must be assessed step-by-step. These two rules support correct decision-making based on two basic principles:

Principle 1. Innovation-based components introduced into the OTLS should interact with standard-based components embedded into that system earlier. If the resulting composition is non-interoperable, then the new component should be excluded, because the target OTLS cannot continue to exist under the formed conditions, or all components should be co-evolved and, as a result, mutually adapted within the system.

Principle 2. If the OTLS is under transformation then all phenomena that occur within the technical system should have requirements changed and as a result, the composition of this OTLS will be modified.

Regarding continuous market changes, OTLS almost constantly is under one of the so-called transitive conditions, described by critical parameters of SC (*Structural Compatibility*) and FI (*Functional Interoperability*). Moreover, the larger the variety of topological compositions is allowable for system structure which is in transitive condition, the more likely the target system is able to adapt to external impacts.

Following the above-mentioned argumentation, the authors propose the five-steps procedure of converting a renewal task into improved transport structure:

Step 1. Identification of OTLS modernization problem. Obsolete elements are labeled, as well as possible solutions are identified.

Step 2. Identification of relations between the OTLS elements. Both direct and indirect interactions, as well as binary operations for heterogeneous coalitions selection are indicated.

Step 3. OTLS hierarchy development, i.e. the OTLS elements composition order. Authors suggest using algebraic topology for innovative coalitions integration into the OTLS structure.

Step 4. Innovative elements implementation. Improvement process combines step-by-step analysis of SC & FC between innovative and obsolete coalitions with accepted or non-accepted changes within the OTSL structure.

Step 5. Approval of the strategy of large-scale commercialization for the upgraded OTSL. This step finishes the procedure and isn't considered in presented paper.

The idea of using the topology as an optimization tool for structural analysis of the artificial systems has been proposed by Kost B. (Kost, 1995). On the other hand, Levin M. (Levin, 2015) has explored the concept of upgraded system structure constructed from previously selected elements. According to mentioned publications, the topology of each OTLS can be described as the skeletal diagram shown in Fig.7

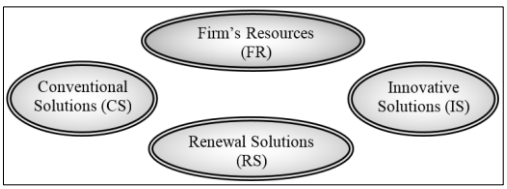


Fig. 7. Basic coalitions of the typical OTLS with uncertain connection state

The topology of each hybrid system is defined on a fixed planar grid with few-components. In our case, OTLS, as multilayer structure, is made from separate coalitions of engineering solutions. Such solutions are grouped into CS (*Conventional Solutions*), IS (*Innovative Solutions*) and RS (*Renewal Solutions*). Considering the OTLS as a hybrid system, we are dealing with the complex problem, because:

- modernisation process is performed on multilayered structure (Fig. 8),
- the success of modernisation process depends on designer's knowledge and skills.

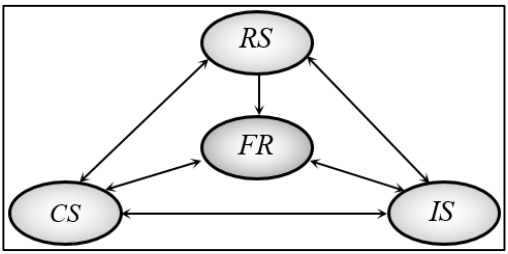


Fig. 8. Basic coalitions of the typical OTLS with certain connection state

Regarding this problems the following situations for OTLS modernization can take place:

1) **Absolutely uncertain situations** when the designer doesn't possess any information about desired innovation. In this case, *CS* is replaced by *RS*. Such solutions are called *BF (Braking Factors)*, and consequently increase the regressive trends in the OTLS development. The *BF* elimination is possible due to the strengthening of the modernising strategies and best transport practices implementation.
2) **Partial uncertain situation** when the designer possesses only partial information about desired innovation. However, final solutions are connected with the high risk and depend on top-management decisions and available assets.
3) **Certain situation** when the designer has complete and reliable information about desired innovation, as well as diffusion of the desired innovation has stable behavior.

3.3. The topological approach to meso-level modernization

Let's assume that typical OTLS is a hypothetical construct that represents a multi-layer complex and consists from conventional component-based coalitions achieving a particular obsolete level. The actual demands imposed by transport and logistics tasks during their performance may be modified by many factors (e.g., the increasing competition, new legal norms, various technical defects, client's wishes) that require the system modernization. Such changes the most commonly appear on the OTLS meso-level. Let's particularize the above-mentioned steps of the OTLS modernization on chosen structure example (Fig. 9):

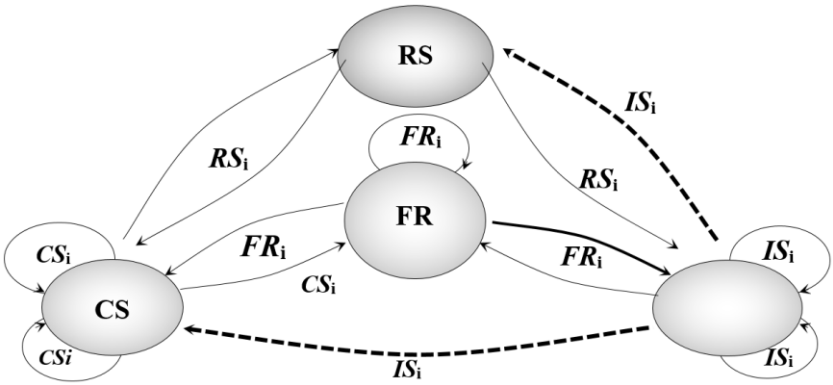


Fig. 9. The topological diagram of typical ways for OTLS modernization

Step 1. Identification of outdated elements and modernization/adaptation activities. The following actions between basic coalitions can be recognized:

A. The CS– coalition:
- *Link* CS_1 – labeling obsolete components;
- *Link* CS_2 – marking interrelated components;
- *Link* CS_3 – evolving standardization norms;
- *Link* CS_4 – reconditioning conventional solutions.

B. The IS – coalition:
- *Link* IS_1– testing before implementation;
- *Link* IS_2 – evolving structural & functional effects;
- *Link* IS_3 – marking PZSI (*Zone of Structural Incompatibility*);
- *Link* IS_4 – marking AFNI (*Area of Functional Non-Interoperability*);
- *Link* IS_5 – elimination of the detected incomplete effects.

C. The RS – coalition:
- *Link* RS_1 – renovating obsolete components;
- *Link* RS_2 – renewing database of innovative solutions;
- *Link* RS_3 – changing of operation mode.

D. The FR (*Firm's Resources*):
- *Link* FR_1 – improving adaptation techniques;
- *Link* FR_2 – permanent control of capital & operating expenditures;
- *Link* FR_3 – decision-making support.

Step 2. Identification of coalitions relations within the OTLS.

Each OTLS could be defined as orderly set coalitions and relationships between them. Then typical OTLS will be described as multilayered topological structure (Fig. 9), which contains:
- three pairs of the inter-coalitions relationships:

$$(CS_i - RS_i);\ (FR_i - CS_i);\ (IS_i - FR_i); \quad (1)$$

- three pairs of the self-oriented relationships:

$$(CS_i);\ (FR_i);\ (IS_i); \quad (2)$$

- two pairs of the single-oriented relationships namely IS_i links.

Usually, the modernisation process is realised via three techniques:

1) Firstly, the technique for creating future-oriented OTLS through the replacing conventional solutions by innovative solutions. This replace is possible as:

1.1) Direct change of conventional solutions by innovative solutions:

$$FR_i + IS_i = OTLS_{i+1} \quad (3)$$

1.2) Indirect change of conservative solutions by new solutions (IS to RS and RS to CS):

$$FR_i + IS_i + RS_i = OTLS_{i+1} \quad (4)$$

2) Secondly, the technique for creating modern OTLS through the replacing conventional solutions by renewal solutions:

$$FR_i + CS_i + RS_i = OTLS_{i+1} \quad (5)$$

3) Finally, the technique for increasing number of conventional solutions through innovative and renewal standardisation:

$$FR_i + RS_i = CS_{i+1} \quad (6)$$

To conclude, decision-maker determinates two groups of modernization procedures:

- the procedures of evolutionary process, mapped as:

$$FR_i + IS_i + OTLS_i = OTLS_{i+1} \quad (7)$$

- the procedures of co-evolutionary process, mapped as:

$$(IS_i - FR_i) + (RS_i - FR_i) + (RS_i - IS_i) = OTLS_{i+1} \quad (8)$$

Step 3. The hierarchy of the upgrading process.
Let's create topological sequence for the possible transformations within the OTLS. For that purpose we use the skeleton diagram (Fig. 9) and apply modernization procedure assuming that this system is upgraded at the component level only. Firstly, such procedures are used for each coalition separately, and secondly - for inter-coalition relations.

Step 4. Innovative elements implementation.
We should consider each vertex of our framework through the sequence of elementary processes describing implementation of innovative elements, under condition of authentic cohesion of the target OTLS (shown in Figure 10).

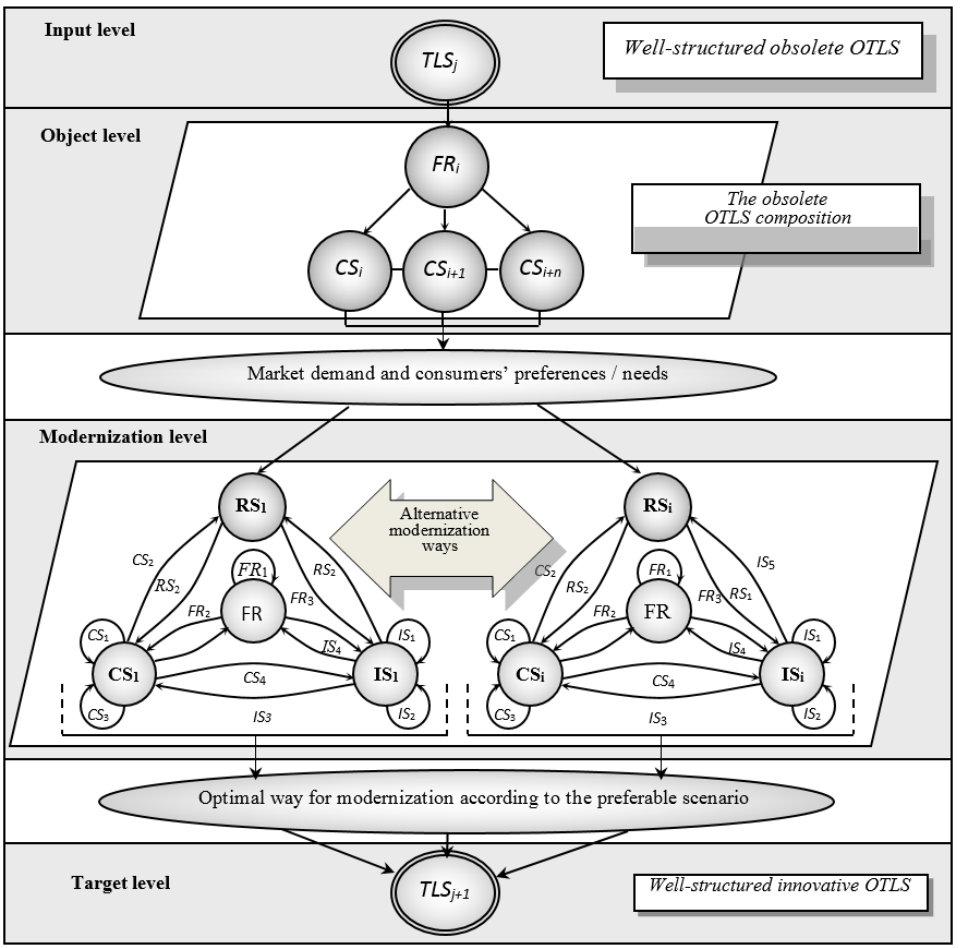


Fig. 10. The conceptual framework for the OTLS modernization sequence

4. Proof - of - Concept study

The studies conducted by authors were divided into three related tasks:

1) Establishment of the state-of-the-art in the extent of SECA requirements;
2) Assessment of success factors, barriers and transferability effects of innovative solutions;
3) Development of the approach to integrate standard and innovative components within the obsolete MPSs (*Marine Propulsion Systems*) under condition of cohesion.

4.1. Explored innovative solutions

Let's consider innovative solutions implementation in shipping industry in viewpoint of modernisation of marine propulsion, cleaner fuels and green technology introduction. There are many factors affecting choice of the suitable ship emissions reduction method, which include ship type, power rating, economic issues, adaptability, and compliance with the current and future emission regulations. The authors have studied the basic ones (Table 2).

4.2. Explored barriers to innovative solutions

In order to foster innovation competition dynamics and attenuate systemic failures, it is important to identify the multiplicity of barriers faced by future-oriented companies According to several authors (Løvdal and Neumann, 2011; Semenov, 2008) barriers for maritime business innovation are identified within the multidimensional framework along the five groups of causes, namely technological, financial, legal, market and management specifics. Furthermore, most of identified barriers emerge or tend to aggravations between shipping and shipbuilding industries. The results of conducted research are shown in Table 3.

The research revealed that:

- technological barriers represent the most numerous group (26.6%);
- financial barriers form the second largest group (25.2%);
- the least onerous barriers are legal barriers (4.8%).

On the other hand:

- fuel–related technology is easy to implement in shipping operations (15.6%);
- maximum level of complication in new technology implementation regards the dual fuel engine technology (31.9%).

Table 2. Chosen examples of innovative solutions

No	Innovative solution	Characteristics
1	Fuel–related technology: *Selective catalytic reduction (SCR)*	SCR is a simple, cost-effective NOx reduction solution. The technology uses a simple chemical reaction to neutralise the NOx in the exhaust. Consequently, a NOx limit of less than 0.5 g/kWh can easily be achieved. The investment costs are between 15 and 70 EUR per kW engine power. It depends on engine size and the number of engines per ship. The running costs are mainly driven by the cost for the urea solution. In general, running and maintenance costs are between 5 and 7 EUR per MWh engine power.
2	Engine tuning technology: *Fuel-efficient engines (FEE)*	FEE technologies implementation can help ship engines potentially reduce emissions by 40%. Such vessel should be equipped with latest energy efficient technologies: an exhaust gas by-pass system, a ballast water treatment system, an electronically-controlled engine that can reduce NOx emissions.
3	Exhaust cleaning technology: *Scrubbers (ECT)*	A scrubber is a system that uses seawater and chemicals to remove sulphur from engine exhaust gas. The scrubber uses a chemical reaction to neutralize the SOx present in the exhaust gas. The price for installing a scrubber on a ship typically ranges from EUR 1 million to EUR 5 million per ship, depending on the size of the vessel. SWECO AB* estimates the market for scrubbers and stated that 350 ships have adopted the technology by January 2015. Scrubbers can be included in new ships or retrofitted into existing vessels.
4	In–engine technology: *Dual fuel engine (DFE)*	Another option for shipping companies trying to reduce their sulphur emissions would be to opt for low-sulphur fuel. Low-sulphur fuels are typically marine fuels with a sulphur content that is much lower than heavy fuel oil, which has a sulphur content up to 4.5%. Therefore, the use of low-sulphur fuel is the best solution as it requires limited initial investment costs.
*SWECO AB (originally "Swedish Consultants") is one of the larger European engineering consulting companies, active in the fields of construction, architecture, and environmental engineering.		

Source: Andreasen and Mayer (2007); Chryssakis et al. (2014); IACCSEA, (2012); Lamas et al. (2013); Seddiek and Elgohary (2014).

Table 3. Barriers to meet SECA requirements

Barriers	Causes	Innovation types (%)				Total (%)
		SCR	ECT	FEE	DFE	
Technological barriers	• Structural incompatibilities, • Functional non- interoperability, • The lack of experience, • Producing incompatibilities.	5.5	6.1	6.9	8.1	**26.6**
Financial barriers	• The lack of financial access, • Poor financial background, • High cost of renovated propulsion system.	2.3	6.8	7.8	8.3	**25.2**
Legal barriers	• The lack of patents portfolio, • Prohibitive legislations, • Strong licensure laws etc.	1.7	0.6	1.1	1.4	**4.8**
Market barriers	• Opposition of competitors, • Limited market capacity, • The lack of demand, etc.	3.3	5.2	6.4	7.6	**22.5**
Management barriers	• The incorrect forecast, • The lack of ambitions, • The resistance to changes, • The lack of skilled human resources.	2.8	4.2	7.4	6.5	**20.9**
Total		**15.6**	**22.9**	**29.6**	**31.9**	**100**

Source: Authors research on the basis: Godfrey (2008); Hölzl and Janger (2011); Roithmayr (2000); Semenov (2008)

4.3. Methodological base for the assessment of innovative solutions compatibility

In order to assess possible effects of two element-based coalitions connection, we propose to determine the three levels of innovative solutions compatibility (Table 4). The interaction of coalitions was investigated distinguishing:

1) Highly compatible coalitions, where two standard coalitions are considered and their interaction contributes to increase of the conventional OTLS efficiency.
2) Coalitions interaction that affects the need of radical changes in the OTLS structure. In this case the standard and innovative coalitions are connected. Heterogeneity of coalitions structure causes insufficient capability of OTLS, therefore structural changes should be made.
3) Incompatible coalitions, where two incoherent coalitions interaction is analyzed.

Each of mentioned group of innovative solutions used in modernisation of MPS was analyzed in detail and the relationships between particular element-based coalitions were investigated.

4.4. The early-stage results

Conducting the research on coalitions fitting the authors received introductory results. The topological approach was used to analyze the retrofitting of MPSs. The space of simulation results was broken down by four subspaces shown in Fig. 11, wherein:

- two subspaces (matched as 2, in Fig.11) are covered by ZSI (*Zone of Structural Incompatibility*), where so-called incomplete effects of retrofitting process are located;
- one subspace (matched as 1, in Fig.11) is covered by ZFC (*Zone of Full Compatibility*), where on-target effects of retrofitting process are set;
- one subspace (matched as 3, in Fig.11) is covered by AFNI (*Area of Functional Non-Interoperability*), where off-target effects of retrofitting process take place.

Analyzing the research results (Fig. 11) it can be stated that each modernisation iteration have different innovative effects. On-target effects are achieved within ZFC of coalitions fitting, while off-target effects are specific for AFNI.

Table 4. Domains describing the levels of innovative solutions compatibility

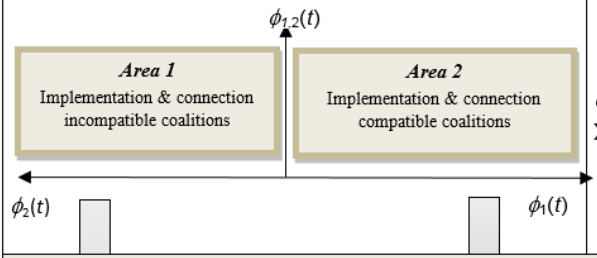
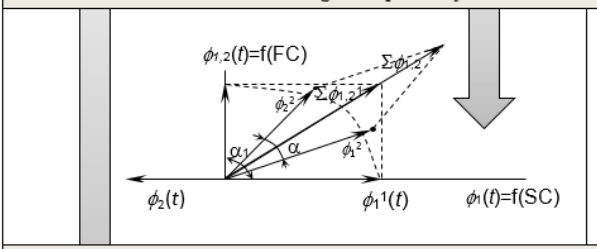
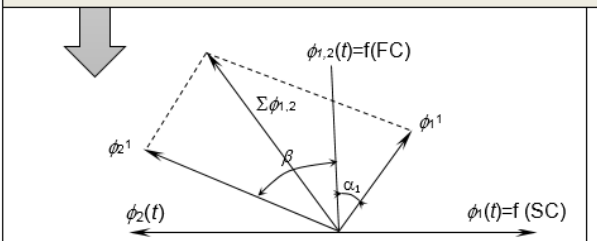
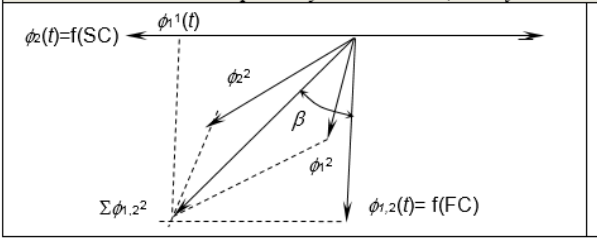
Graphical illustration	Comment
	1. *Task:* two coalitions connection 2. *Nomenclature:* $\phi_1(t)$ = potential of the 1-st coalition, $\phi_2(t)$ = potential of the 2-nd coalition, $\Sigma\phi_{1,2}$ = innovative ship potential, α = compatibility level, β = incompatibility level
1. High compatibility level of the two coalitions	
	1. Modernization stage: Implementation and connection of two compatible coalitions 2. Evaluation: $\beta = 0$ 3. Conclusion: On-target effect
2. Partial compatibility level - PZSI (*Partial Zone of Structural Incompatibility*)	
	1. Modernization stage: Implementation of innovative coalition connected with standard coalition 2. Evaluation: $\beta > \alpha$ 3. Conclusion: Incomplete effect
3. Incompatibility level - AFNI (*Area of Functional Non-Interoperability*)	
	1. Modernization stage: Implementation of innovative coalition connected with standard coalition 2. Evaluation: $\alpha = 0$ 3. Conclusion: Off-target effect

Figure 11 presents one of the tested sequences of the modernisation process for the obsolete MPSs. The conducted research in this subject area revealed incompatible innovative and standard coalitions, where 18% of structural compatibility and 12% of functional interoperability were indicated.

The research results were also grouped into six categories determined by various factors, e.g. kinds of innovative activities and expected effects, compatibility of innovative and outdated coalitions, etc. (Fig.12).

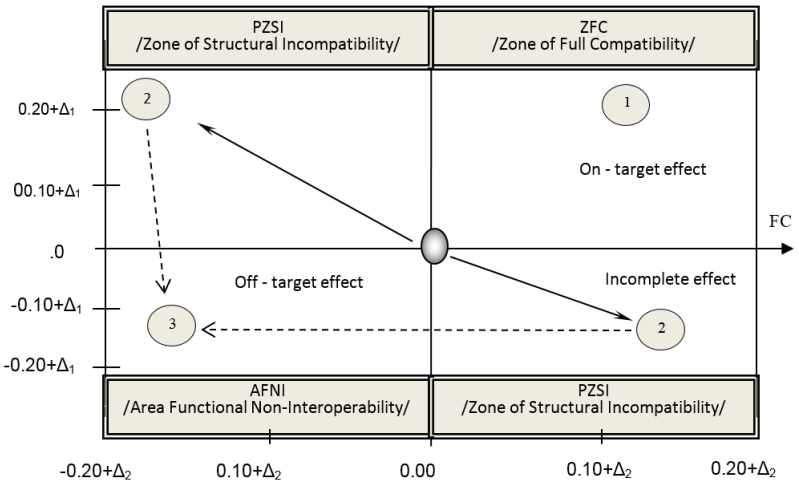


Fig. 11. The subspaces of innovative effects (research fragment)

Each of the selected solutions for modernisation of marine propulsion should be based on relevant and actual information about goals and different advantages/disadvantages of particular solutions (Tables 1 and 2).

4.5. Results of compatibility and interoperability analysis

The compatibility analysis is required to assess the feasibility of using the competing innovative coalitions. Specific details of the implementation options regarding innovative marine propulsion systems are currently not validated. Therefore, our research considers only compatibility of the innovative and outdate coalitions.

The overall study addresses two aspects of current MPSs. In introduction part, SCR, FEE, ECT and DFE are considered. The first aspect relates to current MPSs as obsolete and aims to define the appropriate spectrum of perspective innovation solutions. The second one considers the current MPSs as interferers and aims to evaluate the impact of innovative coalition to current coalitions within MPSs.

The criteria identified in this paper are based on the current expected operating conditions of the new MPSs, including protection and susceptibility criteria for the maritime industry modernization. They regard such issues as i.a. emission limits for sulphur content (no more than 0.10% from 1st January 2015, against the limit of 1.00% until 31 December 2014), as well as expectations of the structural and functional parameters.

New MPS must be tuned to tolerate interference from other ship systems to operate in different sea conditions. Such systems should have compliance at levels exceeding the 99.9% for functional parameters, and 100% for the structural parameters. In order to choose the best upgraded solution the following steps may be used:

1) The objectivities choice and analysis. It should be based on the 2030 Agenda for Sustainable Development (introduced by the United Nations in September 2015 (IMO, 2016)), i.a. the goal 9 to "build resilient infrastructure, promote inclusive and sustainable industrialization and foster innovation". The objectivities spectrum should include:
   - *structural compatibility* - coincides with the objective of the structure integrity;
   - *functional interoperability* - coincides with the ability of few coalitions to operate effectively and efficiently together;
   - *capital expenditure* - coincides with costs to acquire or upgrade productive assets;
   - *operational expenditure* - coincides with lifecycle cost;
   - *maintenance cost* - the costs associated with keeping propulsion system in good condition by regularly checking it and repairing it when necessary.

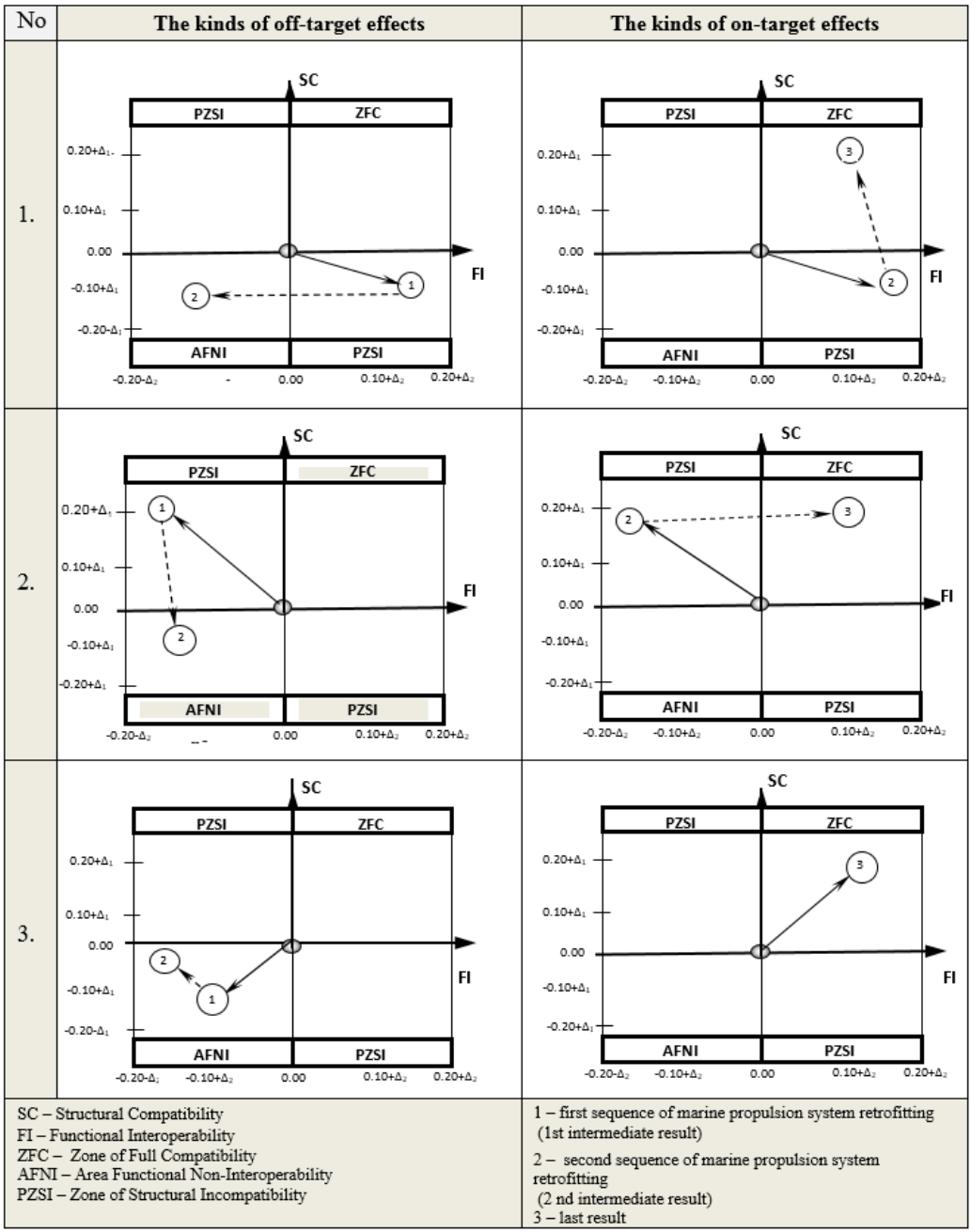


Fig. 12. Possible variants of predicted and unforeseen effects of innovative solutions implementation

2) Realization of two-step upgrading and assessment procedures:
   - Using the set of six categories proposed earlier (Fig. 12) to analyse the possible modernisation situations. We propose to assess the upgrade effectiveness of modernisation according to structural compatibility and functional interoperability.
   - To carry out next phase of upgrading and assessment of the MPS modernization according to selected measures, including "*Capital expenditure", "Operational expenditure", "Maintenance cost*".

3) Choice of the best upgraded solution.

The intermediate results based on the simulation tests of the structural compatibility and functional interoperability of MPSs after implementation of innovative solutions are given in Fig. 13. Assessment of the upgrade effectiveness of modernisation according to structural compatibility and functional interoperability give us only intermediate results.

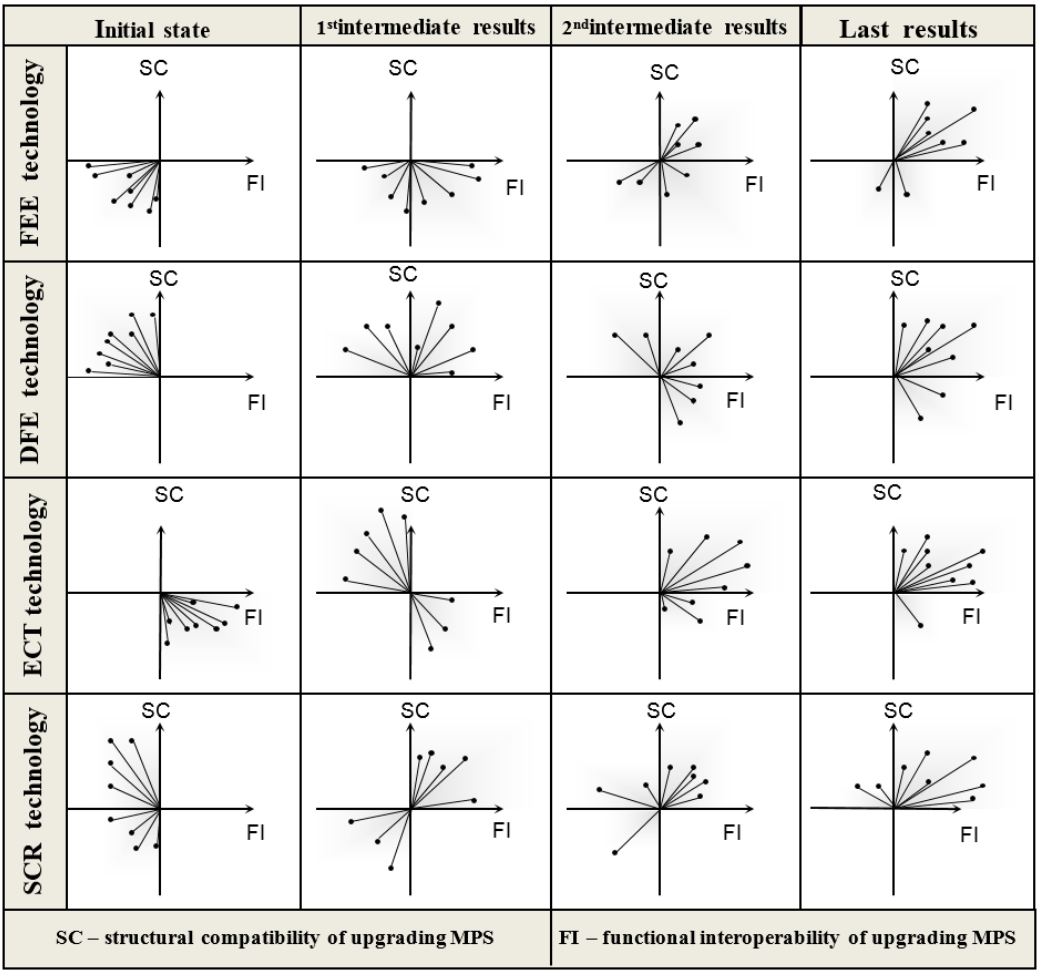


Fig. 13. Boxplots intermediate results of first step modernization for marine propulsion system

Below there are the results of the comparative analysis of four investigated innovative solutions according to selected financial measures. The calculations were carried out in accordance with Ventura M. (2017), where the following three measures were investigated:

1) Capital expenditure:

$$C_m = 1.6 \times \left({}^{P_B}/_{100}\right)^{0.82} \times m_M + CF_M\ [USD] \quad (9)$$

where:
P_B – propulsive power (MCR) [kW],
m_M – unit cost of the machinery [USD /kW],
CF_M– installation and alignment cost of the machinery [USD].

2) Maintenance & repair cost:

$$C_{M\&R} = k_1 \times C_0 + k_2 \times P_{MCR}^{0.66} \quad [USD/year] \quad (10)$$

where:
C_0– cost of the ship [USD],
P_{MCR}– propulsive power [hp],
k_1, k_2– coefficients that depend of the type of propulsion plant ($k_1 = 0{,}0035;\ k_2 = 125$) .

3) Operational expenditure:

$$C_{SUP} = k_1 \times N + k_2 (Lpp \times B \times T)^{0{,}25} + k_3 \times P_{MCR}^{0.7} \quad [USD/year] \quad (11)$$

with:
$Lpp \times B \times T$– cubic number [m^3],
P_{MCR} – propulsive power [hp],
N- crew number [persons],
k_3= 150 (steam turbine),
k_3= 250 (diesel engine, 4 stroke),
k_3= 200 (diesel engine, 2 stroke),
k_2= 4,000 (freight ship),
k_2= 5,000 (tanker),
k_1=3,500.

The results of comparative analysis as three vertical bar charts are given on Fig.14.

5. Summary and Outlook

Innovation activity has become a key factor for gaining competitive advantages on the transportation and logistics market. Transport industry developed and is developing a wide range of innovative concepts to make transport systems more efficient and competitive. Such activities are focused on management of thematically connected innovative projects.
The authors studied a wide range of innovative solutions for cleaner inland and short sea shipping including fuel–related technology, engine tuning technology, exhaust cleaning technology and dual fuel engine. Research was based on expert groups interviews, as well as desk research of good practice examples. During the research it was found that desirable and undesirable retrofitting effects take place and despite of significant advancement there are numerous barriers for the large-scale implementation of achieved innovative solutions.
Explored aspects are dealing with structural and functional compatibilities required to be successful in ships' modernisation to meet SECA requirements now and in the future. The results show that fuel-related technology and water scrubber systems installed on-board ships are the good methods from the environmental viewpoint. Application of one of them depends on some conditions such as i.a. required emission reduction percentage.

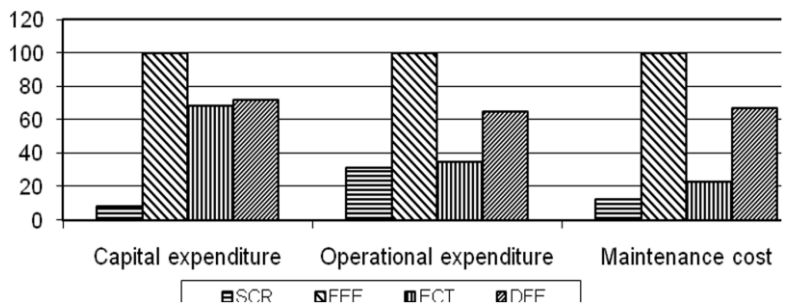


Fig. 14. Comparative profile analysis of evaluated results
Source: Authors research on the basis based on the information contained in the tables 2 and 3.

In authors' option, the dual fuel engine is very competitive technology in long-term perspective when the alternative fuels will be convenient from the market point of view and financial issues.
Achieved results shown that the application of the topological-oriented approach for analysis of the OTLS modernization process is correct. The obtained conclusions are preliminary and will be a subject to further research.

References

[1] AHO, E., CORNU, J., GEORGHIOU, L., and SUBIRA, A., 2006. Creating an Innovative Europe. Report of the Independent Expert Group on R&D and Innovation. Brussels: European Commission.
[2] ANDREASEN, A. and MAYER, S. 2007. Use of Seawater Scrubbing for SO2 Removal from Marine Engine Exhaust Gas. *Energy Fuels*, 21 (6), 3274–3279.
[3] ARM BADR-EL-DIN, A, 2013. Object-Oriented in Organization Management: Organic Organization. *International Journal of Digital Information and Wireless Communications*, 3(4), 440-450.
[4] AZIZ, H., and de KEIJZER, B., 2011. Complexity of coalition structure generation. In 10th International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS), 191–198.
[5] BARAS, J. S., 2011. Cooperative networked systems: Multiple graphs, coalitional games, new probabilistic models. 19th Mediterranean Conference on Control and Automation, Aquis Corfu Holiday Palace, Corfu, Greece June 20-23, 798-800.
[6] BHALLA N., IPPARTHI D., KLEMP E., and DORIGO M., 2014. A Geometrical Approach to the Incompatible Substructure Problem in Parallel Self-Assembly. In: Parallel Problem Solving from Nature – PPSN XIII, series Lecture Notes in Computer Science, Vol. 8672, Springer International Publishing, 751-760.
[7] BOLT, J., and van ZANDEN, J. L., 2014. The Maddison Project: Collaborative Research on Historical National Accounts. *The Economic History Review*, 67(3), 627-651.
[8] Bosch website: http://www.bosch.com.br/content/language1/html/734_4418.htm, access: 02.2017.
[9] CHAMINADE, C., LUNDVALL, B.-A., VANG-LAURIDSEN J., and JOSEPH, K.J, Innovation policies for development: towards a systemic experimentation based approach. CIRCLE Electronic Working Paper Series 2010/01, http://www.circle.lu.se.
[10] CHATAWAY, J., HANLIN, R., and KAPLINSKY, R., 2014. Inclusive Innovation: An Architecture for Policy Development. *Innovation and Development*, 1(4), 33-54.
[11] CHEN, B., & CHENG H.H., 2010. A Review of the Applications of Agent Technology in Traffic and Transportation Systems. *IEEE Transactions on Intelligent Transportation Systems*, 11(2), 485-497.
[12] CHEN, Y.M., and WANG, B.-Y., 2009. Towards Participatory Design of Multi-agent Approach to Transport Demands. *IJCSI International Journal of Computer Science Issues*, 4(1), 10-15.
[13] CHRYSSAKIS Ch., BALLAND O., TVETE H. A., and BRANDSÆTER A., 2014. Alternative fuels for shipping. Position Paper 17–2014, DNV& GL, Norway.
[14] CRESPI, V., GALSTYAN, A. and LERMAN, K., 2008. Top-down vs bottom-up methodologies in multi-agent system design. *Autonomous Robots*, 24(3), 303-313.
[15] DAVIDSSON, P., HENESEY, L., RAMSTEDT, L., and TORNQUIST, J., 2005. An analysis of agent-based approaches to transport logistics, *Transportation Research Part C*, 13, 255-271.
[16] DOSI, G., NELSON, R. R. and WINTER, S. 2000. The Nature and Dynamics of Organizational Capabilities. Oxford: Oxford University Press
[17] GODFREY, N., 2008. Why is competition important for growth and poverty reduction? Department for International Development, OECD Global forum on International Investment, London, March 2008.
[18] GRAUDINA V., and GRUNDSPENKIS J., 2005. Technologies and Multi-Agent System Architectures for Transportation and Logistics Support: An Overview. International

Conference on Computer Systems and Technologies - CompSysTech', IIIA.6, 1-6.
[19] HÖLZL, W., and JANGER, J., 2011. Innovation barriers across firm types and countries, the DIME Final Conference, 6-8 April, Maastricht.
[20] HUBKA, V., and EDER, E. W., 1988. Theory of Technical Systems: A Total Concept Theory for Engineering Design. Springer, 249.
[21] IACCSEA: The Technological and Economic Viability of Selective Catalytic Reduction for Ships, December 2012 London, UK.
[22] IMO, 2016. Outcomes of the United Nations Climate Change Conferences held in Bonn in June, August and October 2015 and Paris in December 2015. Note by the Secretariat. MEPC 69/7. London
[23] JACYNA-GOŁDA, I., 2015. Decision-making model for supporting supply chain efficiency evaluation. *Archives of Transport*, 33(1), 17-31.
[24] JUMAN, Z. A. M. S., HOQUE, M. A., and BUHARI, M. I., 2013. A study of transportation problem and use of object oriented programming. 3rd International Conference on Applied Mathematics and Pharmaceutical Sciences (ICAMPS'2013), April 29-30, 2013 Singapore, 353-354.
[25] KERBACHE, L., and MACCREGOR SMITH J., 2004. Queuing Networks and the Topological Design of Supply Chain Systems. International *Journal of Production Economics*, 91, 251-272.
[26] KONINGS, J. W., PRIEMUS, H., and NIJKAMP, P., 2005. The Future of Automated Freight Transport: Concepts, Design, and Implementation. Cheltenham, UK: Edward Elgar, 85.
[27] KOST, B., 1995. Evolution Strategies in Structural Topology Optimization of Trusses. Computing in Civil and Building Engineering. Proceedings of the 6th Intern. Conf., Rotterdam, 675-681.
[28] LAMAS M. I., RODRIGUEZ C. G., RODRIGUEZ J. D., and TELMO J., 2013. Internal modifications to reduce pollutant emissions from marine engines. *International Journal of Naval Architecture and Ocean Engineering*, 5(4), 493-501.
[29] LEVIN, M. Sh., 2015.Modular System Design and Evaluation. Springer, 437.
[30] LIN, J.-Ch. 2011. Various Approaches for Systems Analysis and Design. University of Missouri, St. Louis, available at: http://www.umsl.edu/~sauterv/analysis/termpapers/f11/jia.html.
[31] LØVDAL N., and NEUMANN F., 2011. Internationalization as a strategy to overcome industry barriers - An assessment of the marine energy industry. *Energy Policy*, 39(3), 1093–1100.
[32] Low - cost smart shipping. Maritime industry IoT development. London. 2017, p. 5.
[33] MAURO, N. D., BASILE, T. M. A., FERILLI, S., and ESPOSITO, F., 2010. Coalition structure generation with grasp.14th Int. Conf. on Artificial Intelligence: Methodology, Systems, and Applications (AIMSA), 111-120.
[34] MODELEWSKI, K., and SIERGIEJCZYK, M., 2013. Application of Multi-agent systems in transportation, *Prace Naukowe Politechniki Warszawskiej*, 100, 145-152.
[35] RAHWAN, T., RAMCHURN, S. D., GIOVANNUCCI, A., and JENNINGS, N. R., 2009. An anytime algorithm for optimal coalition structure generation. *Journal of Artificial Intelligence Research (JAIR)*, 34, 521-567.
[36] ROCHA A., RIBEIRO, L., and BARATA, J., 2014. A Multi Agent Architecture to Support Self-organizing Material Handling, in (Eds. Camarinha-Matos L. M. et al.) Technological Innovation for Collective Awareness Systems, Vol. 423 of the series IFIP Advances in Information and Communication Technology, Springer, 93-100.
[37] ROITHMAYR, D., 2000. Barriers to Entry: A Market Lock-In Model of Discrimination, *Virginia Law Review*, 86(4), 727-799.
[38] SCHENK, N. J., MOLL, H. C., and SCHOOT UITERKAMP, A. J. M., 2007. Meso-level analysis, the missing link in energy strategies. *Energy Policy*, 35(3), 1505-1516.
[39] SCHUMPETER, J. A. 1950. Capitalism, socialism, and democracy. 3d ed. New York: Harper and Row.
[40] SEDDIEK I. S., and ELGOHARY M. M., 2014. Eco-friendly selection of ship emissions

reduction strategies with emphasis on SOx and NOx emissions. *International Journal of Naval Architecture and Ocean Engineering*, 6(3), 737-748.

[41] SEMENOV, I. N., 2008. The multidimensional approach to marine industry development. Part I. Obstacles and willingness to the EU marine industry reengineering. *Polish Maritime Research*, 3(57), Vol. 15, 77-85.

[42] SEMENOV, I.N., 2006. Co-evolution approach to management by the transport networks' innovative transformations. Part 1. The basic problems and trends innovative transformations. Archives of Transport, 18(1), 49-70.

[43] United Nations Industrial Development Organization, 2015. Industrial Development Report 2016. The Role of Technology and Innovation in Inclusive and Sustainable Industrial Development. Vienna.

[44] VENTURA M., 2017. *Costs Estimate*. Presentation. https://www.coursehero.com.

[45] VOICE T., POLUKAROV M., and JENNINGS N.R., 2012. Coalition Structure Generation over Graphs. *Journal of Artificial Intelligence Research*, 45, 165-196.

ARCHIVES OF TRANSPORT www.czasopisma.pan.pl PAN www.journals.pan.pl
Volume 43, Issue 3, 2017
ISSN (print): *0866-9546*
e-ISSN (online): *2300-8830*
DOI: *10.5604/01.3001.0010.4230*

DEVELOPMENT OF A MATHEMATICAL MODEL OF THE GENERALIZED DIAGNOSTIC INDICATOR ON THE BASIS OF FULL FACTORIAL EXPERIMENT

Viktor G. Sychenko[1], Dmytro V. Mironov[2]

[1,2] Dnipropetrovsk National University of Railway Transport named after academician V. Lazaryan, Faculty "Energy processes management", Department "Intellectual power supply system", Dnipro, Ukraine

[1]e-mail: elpostz@i.ua

[2]e-mail: dmitriy.mironov1991@gmail.com

Abstract: *Purpose. The aim of this work is to develop a mathematical model of the generalized diagnostic indicator of the technical state of traction substations electrical equipment. Methodology. The main tenets of the experiment planning theory, methods of structural-functional and multi-factor analysis, methods of mathematical and numerical modeling have been used to solve the set tasks. Results. To obtain the mathematical model of the generalized diagnostic indicator, a full factorial experiment for DC circuit breaker have been conducted. The plan of the experiment and factors affecting the change of the unit technical condition have been selected. The regression equation in variables coded values and the polynomial mathematical model of the generalized diagnostic indicator of the circuit breaker technical condition have been obtained. On the basis of regression equation analysis the character of influence of circuit breaker diagnostic indicators values on generalized diagnostic indicator changes has been defined. As a result of repeated performances of the full factorial experiment the mathematical models for other types of traction substations power equipment have been obtained. Originality. An improved theoretical approach to the construction of generalized diagnostic indicators mathematical models for main types of traction substations electric equipment with using the methods of experiments planning theory has been suggested. Practical value. The obtained polynomial mathematical models of the generalized diagnostic indicator D can be used for constructing the automated system of monitoring and forecasting of the traction substations equipment technical condition, which allows improving the performance of processing the diagnostic information and ensuring the accuracy of the diagnosis. Analysing and forecasting the electrical equipment technical condition with the using of mathematical models of generalized diagnostic indicator changes process allows constructing the optimal strategy of maintenance and repair based on the actual technical condition of the electrical equipment. This will reduce material and financial costs of maintenance and repair work as well as the equipment downtime caused by planned inspections and repair improving reliability and uptime of electrical equipment.*

Key words: *electricity, traction substation, maintenance, diagnostics, full factorial experiment, mathematical model, regression equation.*

1. Introduction

The efficiency and reliability of traction substations electrical equipment (EE) depends on its technical condition. Modern EE has a fairly high estimates of reliability (Szeląg, A., 2017), however, in the process of operation under the influence of external conditions and variable modes of operation the initial state of the equipment is continuously deteriorating, which leads to additional energy losses, reduction of operational reliability and the growth rate of equipment and the number of failures. The reliability of EE during the life cycle depends not only on the quality of manufacture, but also on the operating conditions and quality of maintenance and repair. The providing an inspection, verification, regulation and monitoring of the technical condition of EE and using an innovative technologies of diagnostics are the basis of the strategy of improving the technical object operational reliability (Matusevych, O. O. & Mironov, D. V., 2015; Matusevych, O. O. & Sychenko, V. G. & Bialon A., 2016). A new direction in development of technical

maintenance and repair system is the development of approaches based on individual observation and prediction of the real change of the equipment technical condition during operation. It is necessary to develop a means of obtaining diagnostic information as well as mathematical methods and models that take into account the main factors influencing the technical condition of EE (Szubartowski, M., 2013).

As is known, for assessment of technical state the mechanical (vibration), thermal, electrical and other factors are analyzed. These factors have different physico-chemical nature and lead to EE individual properties changes. In this case, the assessment of technical condition of electrical equipment individual properties runs more or less satisfactorily. However, the overall assessment of the equipment technical condition is extremely difficult due to the need comparing the different physical nature factors and the absence the relationship between them, which can be described by the analytical equation. Therefore, it is proved that for adequate evaluation of the quality of power equipments operational indicators it is necessary using the generalized characteristics of their work (Mironov, D. V., 2015). Knowledge of the equipment generalized technical condition allows assessing the reliability of the whole technological complex. Using a generalized diagnostic indicator as a basic EE technical condition evaluation allows to increasing the accuracy of determining the actual equipment technical condition. However, automating the process of EE diagnosis involves the using of mathematical models of the generalized diagnostic indicator of the equipment technical state to improving the performance of the process and ensuring the reliability of the diagnosis. The task of modeling and forecasting changes in EE technical state is multifactorial and nonlinear, which leads to the application of various methods of its solution. One way of solving this problem is the use of mathematical methods of the experiment planning theory.

To construct a mathematical model of a generalized diagnostic parameter the factors affecting the change of the EE technical condition, the form of the mathematical model and the optimum plan of the experiment are chosen (Adler, Yu. P. & Markova, E. V. & Granovskiy, Yu. V., 1976; Gard, M. & Levinson, S. J. & Ferraro, S. B. & Jimenez, J. A., 2012).

Factors that are used in the planning of the experiment should satisfy the following requirements (Adler, Yu. P. & Markova, E. V. & Granovskiy, Yu. V., 1976):

1) *Controllability*.
2) *Operationally*.
3) *Measurement accuracy*.
4) *Uniqueness*.

During the planning of the experiment the several factors are changed.at the same time. Therefore it is very important to formulate the requirements, which apply to a combination of factors. First of all the requirement of compatibility is determined. Factors compatibility means that all of their combinations are feasible and safe. During the planning of the experiment the factors independence is very important, i.e., the possibility of establishing a factor at any level, regardless of other factors levels.

The next step is to define the primary level and range of factors change and their variation on two levels. In this case, if the number of factors is known, we can immediately find the number of experiments required to implement all possible combinations of factors levels:

$$N = 2^k \tag{1}$$

where:

- N is the number of experiments,
- k is a number of factors,
- 2 is the number of levels.

To simplify the notation of the experiment conditions and processing the experimental data, the scales on the axes are selected so that the upper level corresponds to +1, lower to -1, and the main to zero. For factors with a continuous determination region, this can always be done using the transform

$$x_j = \frac{\tilde{x}_j - \tilde{x}_{j0}}{I_j} \tag{2}$$

where:

- x_j is the coded value of the factor,
- $\tilde{x}_j$ is the natural value of the factor,
- $\tilde{x}_{j0}$ is the natural value of the basic level,

- I_j is the range of variation,
- j is the number of the factor.

For qualitative factors with two levels, one level is denoted by +1 and the other is -1. The order of the levels doesn't matter.

Simplicity and adequacy are the main requirements in determining the form of mathematical models. It is believed that algebraic polynomials are the simplest models. Polynomials are linear relative to the unknown parameters of the model, which greatly simplifies the processing of the experiment results (Adler, Yu. P. & Markova, E. V. & Granovskiy, Yu. V., 1976). Therefore, as a mathematical model the algebraic polynomials has been chosen in this work. During the plan selection the optimality criteria and the number of experiments primarily is taken into account (Gard, M. & Levinson, S. J. & Ferraro, S. B. & Jimenez, J. A., 2012; Mills, K. L. & Filliben, J. J. & Haines, A. L., 2015). In our case, it is clear that the required plan should be two-level (because we are interested in linear model), orthogonal and rotatably. Orthogonality allows to move along the gradient is proportional to the coefficients of the linear model and independently to interpret the effects. Rotatability provides assured equality of prediction variances with the motion in any direction from the center of the experiment. All these requirements are satisfied by a full factor experiment type 2^n (Nalimov, V. V., 1965).

The matrix of the experiment planning must satisfy the following requirements:

1) The symmetry relative to the center of the experiment

$$\sum_{i=1}^{N} x_{ji} = 0, \tag{3}$$

where:
- j is the number of the factor,
- N is the number of experiments,
- j = 1, 2, ... k.

2) The normalization condition

$$\sum_{i=1}^{N} x_{ji}^2 = N. \tag{4}$$

This is a consequence of the fact that the values of the factors in the matrix are set to +1 and -1.

3) Orthogonality of the planning matrix

$$\sum_{i=1}^{N} x_{ij} \cdot x_{ui} = 0, \ j \neq u, \ j,u = 0,1,2,...,k. \tag{5}$$

4) Rotatability. Point in the planning matrix are chosen so that the accuracy of the optimization parameter values prediction is the same at equal distances from the center of the experiment and does not depend on the direction.

The results of the experiment depending on the setting of factors at a given level are recorded in the last column of the matrix of full factorial experiment. Then the results of the experiment are processed:

1) The calculation of the reproducibility variance.

$$s_{\{y\}}^2 = \frac{\sum_{1}^{N} (y_i - \bar{y})^2}{N-1}, \tag{6}$$

where:
- $\bar{y}$ is the expected value of the response function,
- $N-1$ is the number of freedom levels.

2) Determination of regression coefficients.

$$b_j = \frac{\sum_{i=1}^{N} y_i \cdot x_{ji}}{N}, \ b_0 = \bar{y}. \tag{7}$$

3) Testing the significance of regression coefficients.

$$s_{\{b_j\}}^2 = \frac{s_{\{y\}}^2}{N}, \ \Delta b_j = \pm t \cdot s_{\{b_j\}}, \tag{8}$$

where:
- t is the table value of Student's criterion with the number of freedom levels, which was determined by $s_{\{y\}}^2$ and the chosen significance level (in our case for 0.05).

The coefficient is significant if its absolute value is greater than or equal to Δb_j.

4) Checking adequacy of the model.

$$F = \frac{s_{ad}^2}{s_{\{y\}}^2}, \quad (9)$$

where:

- F is Fisher's criterion. In this case $F_{calc} < F_{tabl}$; f is the number of freedom levels, $f = N - (k+1)$; s_{ad}^2 is the dispersion of adequacy:

$$s_{ad}^2 = \frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{f}, \quad (10)$$

where:

- $\hat{y}_i$ is the response function calculated by the regression equation;
- y_i is the response function obtained in the experiment.

The result of the full factorial experiment is written as a polynomial regression equation (Alzoubi, K. & Lu, S. & Sammakia, B. & Poliks M., 2011). The last step in the experiment is the interpretation of results and decision making after constructing a model. The task of interpreting is the fairly complex. The degree of influence of each factor on the optimization parameter is set (Chen, F. & Ma, X. & Zhao, Y. & Zou, J., 2011). In some tasks the construction of the regression equation for the natural values of the factors is required. The equation for the natural variables can be obtained using the transition formula (2). This eliminates the interpretation possibility and the influence of factors on the values and signs of the regression coefficients. However, the resulting equation can be used to predict the optimization criterion changes.

If the linear model is inadequate, to obtain adequate models the following methods are used: changing the intervals of factors variation, the center plan transfer, the plan completion (Bondar', A. G. & Statyukha, G. A. & Potyazhenko, I. A., 1980; Asaturjan, V. I., 1983). The latter method is associated with the transition to the orthogonal central composite experiment plan.

2. The construction of a mathematical model

To obtain the mathematical model of the generalized diagnostic indicator, we have conducted a full factorial experiment for DC circuit breaker in this work. A generalized diagnostic indicator *D(t)* has been used as the response function (Mironov, D. V., 2015). According to (The Ministry Of Infrastructure, 2008; Kuznecov, V. G. & Galkin, O. G. & Efimov, O. V. & Matusevych, O. O., 2009) the main diagnostic indicators and their changes limits for obtaining D have been identified (table 1).

Table 1. Range of diagnostic indicators

Parameter	The main level	The upper limit	The lower limit
Main contacts press, kgp	30	33	27
The spring tension, kgp	35	38	32
The number of outages, th.	40	60	20
The area of adjoining the main contacts, %	85	94	76
The failure of the main contact, mm	1,125	1,515	0,735

Considering the optimality criteria and the number of factors we have selected the full factorial experiment plan type 2^5 of first order (Nalimov, V. V. & Golikova, T. I., 1980) and composed the full factorial experiment planning matrix (table 2). According to the expression (1) the required number of experiments has been defined ($N = 32$).In this matrix the values of the factors is used in all possible combinations. The first column corresponds to the coefficients of the model constant term, columns from 2nd to 6th correspond to the factors value and the 7-th column is the value of the system response.

Table 2. A fragment of the full factorial experiment planning matrix

The experiment number	x0	x1	x2	x3	x4	x5	*D*
1	1	34	43	40	97	2,5	0.638
2	1	34	47	40	97	2,5	0.601
3	1	38	43	40	97	2,5	0.587
4	1	38	47	40	97	2,5	0.553

To simplify the experiment and processing the experimental data, the transformation from the factors natural values to the coded values has been performed by the expression (2). The planning matrix of the full factorial experiment with the columns of the factors interaction (columns 7 – 32) has been completed (table 3).
The constructed matrix satisfies the requirements of (3) – (5). We have been written the model of the generalized diagnostic indicator in the form of the regression equation:

$$\begin{aligned} D = {} & b_0 + b_1 \cdot x_1 + b_2 \cdot x_2 + b_3 \cdot x_3 + b_4 \cdot x_4 + \\ & + b_5 \cdot x_5 + b_6 \cdot x_1 \cdot x_2 + b_7 \cdot x_1 \cdot x_3 + b_8 \cdot x_1 \cdot x_4 + \\ & + b_9 \cdot x_1 \cdot x_5 + b_{10} \cdot x_2 \cdot x_3 + b_{11} \cdot x_2 \cdot x_4 + b_{12} \cdot x_2 \cdot x_5 + \\ & + b_{13} \cdot x_3 \cdot x_4 + b_{14} \cdot x_3 \cdot x_5 + b_{15} \cdot x_4 \cdot x_5 + \\ & + b_{16} \cdot x_1 \cdot x_2 \cdot x_3 + b_{17} \cdot x_1 \cdot x_2 \cdot x_4 + b_{18} \cdot x_1 \cdot x_2 \cdot x_5 + \\ & + b_{19} \cdot x_1 \cdot x_3 \cdot x_4 + b_{20} \cdot x_1 \cdot x_3 \cdot x_5 + + b_{21} \cdot x_1 \cdot x_4 \cdot x_5 + \\ & + b_{22} \cdot x_2 \cdot x_3 \cdot x_4 + b_{23} \cdot x_2 \cdot x_3 \cdot x_5 + b_{24} \cdot x_2 \cdot x_4 \cdot x_5 + \\ & + b_{25} \cdot x_3 \cdot x_4 \cdot x_5 + b_{26} \cdot x_1 \cdot x_2 \cdot x_3 \cdot x_4 + \\ & + b_{27} \cdot x_1 \cdot x_2 \cdot x_3 \cdot x_5 + b_{28} \cdot x_1 \cdot x_2 \cdot x_4 \cdot x_5 + \\ & + b_{29} \cdot x_2 \cdot x_3 \cdot x_4 \cdot x_5 + b_{30} \cdot x_1 \cdot x_2 \cdot x_3 \cdot x_4 \cdot x_5. \end{aligned} \quad (11)$$

The regression coefficients by the formula (7) have been determined. The statistically significant regression coefficients by the formula (8) have been chosen. As a result of the full factorial experiment the regression equation has been obtained:

$$\begin{aligned} D = {} & 0.5583 + 0.02649 \cdot x_1 + 0.02868 \cdot x_2 - \\ & - 0.02014 \cdot x_3 + 0.02566 \cdot x_4 + 0.02497 \cdot x_5 \end{aligned} \quad (12)$$

The resulting model is checked for adequacy by the Fisher test according to the formula (9). The calculated value of Fisher criterion is $F_{calc} = 2.1$. The table value of the Fisher criterion is $F_{tabl} = 1.8874$ (Adler, Yu. P. & Markova, E. V. & Granovskiy, Yu. V., 1976). Because the condition $F_{calc} < F_{tabl}$ is not met, the model (12) can not be considered adequate.
Considering the obtained result it was decided to finish the plan of full factorial experiment to the orthogonal central compositional plan of second order without loss of information about previous experiences (table 4).

Table 3. A fragment of a coded planning matrix of full factorial experiment

The experiment number	x0	x1	x2	x3	x4	x5	x1x2	x1x3	x1x4	x1x5	x2x3	x2x4	x2x5	x3x4	x3x5	x4x5	x1x2x3	x1x2x4	x1x2x5	x1x3x4	x1x3x5	x1x4x5	x2x3x4	x2x3x5	x2x4x5	x3x4x5	x1x2x3x4	x1x2x3x5	x1x2x4x5	x1x3x4x5	x2x3x4x5	x1x2x3x4x5	D
1	1	1	1	1	1	-1	1	1	1	-1	1	1	-1	1	-1	-1	1	1	-1	1	-1	-1	1	-1	-1	-1	1	-1	-1	-1	-1	-1	0.59216
2	1	1	-1	1	1	-1	-1	1	1	-1	-1	-1	1	1	-1	-1	-1	-1	1	1	-1	-1	-1	1	1	-1	-1	1	1	-1	1	1	0.53429
3	1	-1	1	1	1	-1	-1	-1	-1	1	1	1	-1	1	-1	-1	-1	-1	1	-1	1	1	1	-1	-1	-1	-1	1	1	1	-1	1	0.53851
4	1	-1	-1	1	1	-1	1	-1	-1	1	-1	-1	1	1	-1	-1	1	1	-1	-1	1	1	-1	1	1	-1	1	-1	-1	1	1	-1	0.48589

Table 4. The fragment of the encoded planning matrix of the orthogonal central compositional plan of second order

x0	x1	x2	x3	x4	x5	$x1^2$-α	$x2^2$-α	$x3^2$-α	$x4^2$-α	$x5^2$-α	x1x2	x1x3	x1x4	x1x5	x2x3	x2x4	x2x5	x3x4	x3x5	x4x5	x1x2x3	x1x2x4	x1x2x5	x1x3x4	x1x3x5	x1x4x5	x2x3x4	x2x3x5	x2x4x5	x3x4x5	x1x2x3x4	x1x2x3x5	x1x2x4x5	x1x3x4x5	x2x3x4x5	x1x2x3x4x5	*D*
1	1	1	1	1	-1	0.137	0.137	0.137	0.137	0.137	1	1	1	-1	1	1	-1	1	-1	-1	1	1	-1	1	-1	-1	1	-1	-1	-1	1	-1	-1	-1	-1	0.59216
1	1	-1	1	1	-1	0.137	0.137	0.137	0.137	0.137	-1	1	1	-1	-1	-1	1	1	-1	-1	-1	1	1	-1	1	-1	-1	1	1	-1	-1	1	1	-1	1	0.53429
1	-1	1	1	1	-1	0.137	0.137	0.137	0.137	0.137	-1	-1	-1	1	1	1	-1	1	-1	-1	1	-1	-1	-1	1	1	-1	1	1	-1	-1	1	1	-1	-1	0.53851
1	-1	-1	1	1	-1	0.137	0.137	0.137	0.137	0.137	1	-1	-1	1	-1	-1	1	-1	1	1	-1	1	1	-1	-1	1	1	-1	-1	1	1	-1	-1	1	1	0.48589

After the same transformations and calculations the following regression equation has been obtained:

$$D = 0.556036886 + 0.02709781 \cdot x_1 + 0.02913 \cdot x_2 - \\ -0.02015008 \cdot x_3 + 0.025485 \cdot x_4 + 0.026203 \cdot x_5 + \quad (13) \\ +0.034245 \cdot x_1^2 - 0.03553 \cdot x_3^2 + 0.037449 \cdot x_5^2.$$

We have conducted an audit of the obtained regression model and founded that the regression model is adequate ($F_{calc} = 1{,}535 < F_{tabl} = 1{,}6$) (Adler, Yu. P. & Markova, E. V. & Granovskiy, Yu. V., 1976). The simulation results of the generalized diagnostic indicator shown in fig. 1.

As follows from the analysis of Fig. 1 the value of the generic diagnostic indicator according to the results of the experiment and the regression equation are almost identical ($s_{ad}^2 = 0.002978$). It shows the adequacy of the obtained results.

The analysis of the regression equation (13) allowed defining the character of the diagnostic factors influence on the change of DC circuit breaker technical condition. If you increase the value of factors such as the main contacts press, the spring tension, the area of adjoining the main contacts, the failure of the main contact (the regression coefficients of these factors have a positive sign) and reduce the value of the outages number (the regression coefficient of the factor has a negative sign) the value of the generic diagnostic indicator will increase. If you decrease the values of diagnostic factors such as the main contacts press, the spring tension, the area of adjoining the main contacts, the failure of the main contact and increase the value of the outages number the breaker status will be closer to pre-failure. It will signal the need to remove it from service and execute the maintenance and repair work.

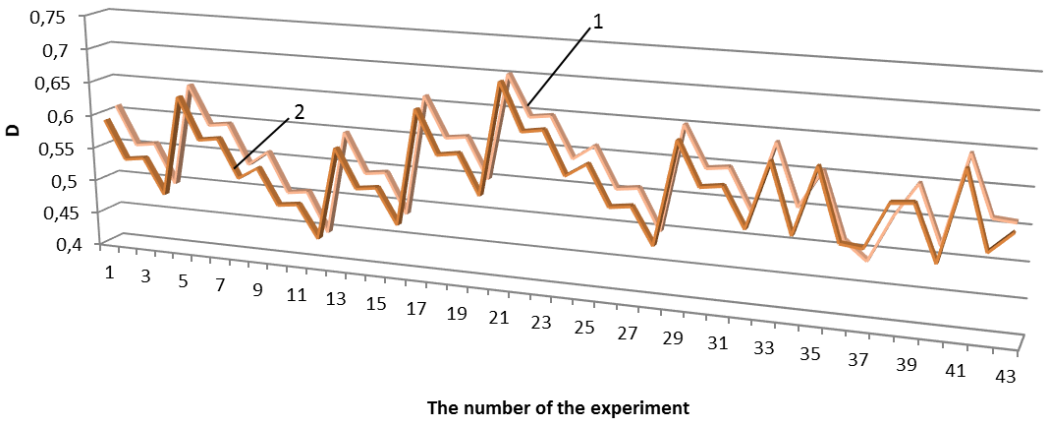


Fig. 1. Generalized diagnostic indicator D :1 – the values of the regression equation, 2 – the values of the experiment results

To obtain the polynomial models of the generalized diagnostic indicator for DC circuit breaker, suitable for use in prediction tasks (Voznesenskij, V. A. & Koval'chuk, A. F., 1978), it is necessary to move from coded to natural variables by the formula (2). After conversion the following expression has been got:

$$D_{DC} = 0.347074 - 0.20529 \cdot x_1 + 0.269718 \cdot x_2 + \\ +0.31173 \cdot x_3 + 0.174552 \cdot x_4 - 0.77776 \cdot x_5 + \\ +0.490774 \cdot x_1^2 - 0.38433 \cdot x_3^2 + 1.135307 \cdot x_5^2. \quad (14)$$

On the basis of this method we have obtained the polynomial model of the generalized diagnostic indicator for other types of traction substations power equipment, namely:

1) Oil circuit breakers ($s_{ao}^2 = 0.002541$)

$$D_{OB} = 0.203348 + 0.333772 \cdot x_1 + 0.327493 \cdot x_2 + \\ +0.593883 \cdot x_3 + 0.226136 \cdot x_4 - \\ -0.49 \cdot x_1^2 - 0.49877 \cdot x_2^2 - 0.29268 \cdot x_3^2. \quad (15)$$

2) Vacuum circuit breakers ($s_{ao}^2 = 0.002254$)

$$D_{VB} = 0.90204 + 0.246541 \cdot x_1 + 0.2479 \cdot x_2 - \\ -0.35461 \cdot x_3 - 0.46874 \cdot x_1^2 - 0.48785 \cdot x_2^2. \quad (16)$$

3) Sulfur hexafluoride circuit breakers ($s_{ao}^2 = 0.002481$)

$$D_{SB} = 0.676946 + 0.31107 \cdot x_1 + 0.091242 \cdot x_2 - \\ -0.29718 \cdot x_3 - 0.31101 \cdot x_2^2. \quad (17)$$

4) Power transformers ($s_{ao}^2 = 0.002256$)

$$D_{PT} = -0.04269 + 0.166659 \cdot x_1 + 0.611565 \cdot x_2 + \\ +0.686812 \cdot x_3 + 0.232403 \cdot x_4 - 0.31107 \cdot x_1^2 - \\ -0.31254 \cdot x_2^2 - 0.37284 \cdot x_3^2. \quad (18)$$

5) Current transformers ($s_{ao}^2 = 0.002815$)

$$D_{CT} = 0.12101 + 0.139881 \cdot x_1 + 0.91763 \cdot x_2 + \\ 0.27498 \cdot x_3 - 0.34259 \cdot x_1^2 - 0.47864 \cdot x_2^2. \quad (19)$$

6) Voltage transformers ($s_{ao}^2 = 0.0025$)

$$D_{VT} = -0.05632 + 0.624604 \cdot x_1 + 0.310983 \cdot x_2 + \\ +0.310983 \cdot x_3 - 0.26262 \cdot x_1^2. \quad (20)$$

The obtained polynomial models of the generalized diagnostic indicator may be used to evaluate and predict the technical condition of the main power equipment of DC traction substations depending on the change of diagnostic parameters. The using of these models for the calculating of the actual technical condition of EE and forecasting its changes for the automation of maintenance and

diagnosis process for traction substations (Mironov, D. V. & Sychenko, V. G. & Matusevych, O. O., 2016) will significantly improve the performance and reliability of the EE monitoring process.

3. Conclusions

In the study it had determined that the change model of the generalized diagnostic indicator D has the form of nonlinear second-order polynomial. It had proved that the using the mathematical methods of the experiment planning theory allows to building a mathematical model of generalized diagnostic indicator D changes the with a sufficiently high accuracy of output result obtaining (s_{ao}^2 for different types of traction substations power equipment varies from 0,002254 to 0,002978). After conducting a full factorial experiment, the regression equation for the traction substations power equipment and the character of the diagnostic parameters influence on the EE technical condition changes have been obtained. To predict changes in the technical condition of power EE the polynomial models in natural values of the diagnostic factors have been obtained. The obtained polynomial mathematical model of the generalized diagnostic indicator D may be used to construct the automated system of monitoring and forecasting of the traction substations equipment technical condition and to improve the performance of the diagnostic information processing and ensure the diagnosis accuracy. The mathematical model of generalized diagnostic indicator can be used in developing of the maintenance and repair optimal strategy based on actual technical condition of EE. This will reduce material and financial costs of maintenance and repair works.

References

[1] ADLER, YU. P. & MARKOVA, E. V. & GRANOVSKIY, YU. V., 1976. *Planning of experiment with the optimal conditions searching*. Moscow: Science.

[2] ALZOUBI, K. & LU, S. & SAMMAKIA, B. & POLIKS M., 2011. Factor Effect Study for the High Cyclic Bending Fatigue of Thin Films on PET Substrate for Flexible Displays Applications. *Journal of Display Technology*, 7(6), 348 -355.

[3] ASATURJAN, V. I., 1983. *The theory of experiment planning*. Moscow: Radio and communication.

[4] BONDAR', A. G. & STATYUKHA, G. A. & POTYAZHENKO, I. A., 1980. *Planning of experiments in the optimization processes of chemical technology*. Kiev: High school.

[5] CHEN, F. & MA, X. & ZHAO, Y. & ZOU, J., 2011. Support Vector Machine Approach for Calculating the AC Resistance of Air-Core Reactor. *IEEE Transactions on Power Delivery*, 26 (4), 2407 – 2415.

[6] GARD, M. & LEVINSON, S. J. & FERRARO, S. B. & JIMENEZ, J. A., 2012. A Factorial Experiment to Characterize the Small-Caliber Launcher. *IEEE Transactions on Plasma Science*, 40 (1), 118-123.

[7] KUZNECOV, V. G. & GALKIN, O. G. & EFIMOV, O. V. & MATUSEVYCH, O. O., 2009. *Reliability and diagnostics of the traction power supply equipment*. Dnipro: Makovec'kij.

[8] MATUSEVYCH, O. O. & MIRONOV, D. V., 2015. Study of the manual power equipment of traction electrification system of the railways. *Science and Transport Progress. Bulletin of Dnipropetrovsk National University of Railway Transport*, 1 (55), 62-77.

[9] MATUSEVYCH, O. O. & SYCHENKO, V. G. & BIALON A., 2016. Continuous improvement of technical servicing and repair system of railway substation on the basis of FMEA methodology. *Technika Transportu Szynowego*, 1-2, 75—79.

[10] MILLS, K. L. & FILLIBEN, J. J. & HAINES, A. L., 2015. Determining Relative Importance and Effective Settings for Genetic Algorithm Control Parameters. *Evolutionary Computation*, 23 (2), 309 – 342.

[11] MIRONOV, D. V. & SYCHENKO, V. G. & MATUSEVYCH, O.O., 2016. Automated system of monitoring and predicting the actual residual life of the traction substations equipment. *Electrical engineering and Electromechanics*, 4 (1), 57 – 63.

[12] MIRONOV, D. V., 2015. Improvement the system of maintenance service and repair for traction substations equipment using the generalized criteria. *Power Engineering:*

Economics, Technique, Ecology, 3 (41), 107-116.

[13] NALIMOV, V. V. & GOLIKOVA, T. I., 1980. *Logical basis of experiment planning*. Moscow: Metallurgical.

[14] NALIMOV, V. V., 1965. *Statistical methods of extreme experiments planning*. Moscow: Science.

[15] SZELĄG, A., 2017. Electrical Power Infrastructure for Modern Rolling Stock with Regard to the Railway in Poland. *Archives of Transport*, 42 (2), 75-83.

[16] SZUBARTOWSKI, M., 2013. Semi-Markov Model of the Operation and Maintenance Process of City Buses. *The Archives of Transport*, 1-2, 109-116.

[17] THE MINISTRY OF INFRASTRUCTURE, 2008. Manual on maintenance and repair of the traction substations equipment of electrified Railways.

[18] VOZNESENSKIJ, V. A. & KOVAL'CHUK, A. F., 1978. *Decisions on statistical models*. Moscow: Statistics.