The Journal of Transport and Land Use http://jtlu.org
Vol. 11 No. 1 [2018] pp. 593–632

JTLU

Theoretical substantiation of trip length distribution for home-based work trips in urban transit systems

Peter Horbachov
Kharkiv National Automobile and Highway University
gorbachev_pf@mail.ru

Stanislav Svichynskyi
Kharkiv National Automobile and Highway University
stas_svichinsky@ukr.net

Abstract: Modern approaches to the modeling of transport demand imply the use of calibration procedures during the origin-destination (O-D) matrix estimation or transit assignment. These procedures lead to misrepresenting generated and attracted trips or changing the trip length distribution (TLD). It means that the methods of transport planning can be improved by means of determination, validation and implementation of the TLD to calculate the O-D matrix. The analysis of research results in the field of mass transit reveals an explicit similarity between TLD in different cities and the gamma distribution. It points to general regularities in various systems of mass transit that lead to the similarity in TLD. The regularities are determined by studying the spatial distribution of mass transit stops, which are considered trip origins and destinations. The experimental research was conducted in 10 Ukrainian cities using probability theory methods.

Keywords: trip length distribution, urban transit, trip attractor, transit stop, transport demand, Ukraine

Article history:
Received: June 21, 2016
Received in revised form: October 13, 2017
Accepted: January 29, 2018
Available online: August 15, 2018
Data availability: http://doi.org/10.3886/E100393V1

1 Introduction

In recent years, the Eastern European countries have widely used the methods of mathematical modeling during the planning of transit system operations in cities and regions. The most important phase in such planning is to form a demand model, specifically, an O-D matrix. Modern transport science enables us to estimate the total number of trips generated and attracted by transportation zones (total trip ends) with considerable accuracy, but real directions of passenger trips are not a well-studied. At the same time, the quality of demand modeling is evaluated by means of a comparison of the values of estimated passenger flows with real ones. The difference between them is eliminated in two ways: either by correcting the total trip ends in transportation zones or by changing the deterrence function in the model of trip distribution.

The first approach appears rather doubtful as it distorts real data, namely total trip ends, to reach equivalence with other real data—passenger flows. It would actually be more correct to double-check

http://dx.doi.org/10.5198/jtlu.2018.916
ISSN: 1938-7849 | Licensed under the Creative Commons Attribution – Noncommercial License 4.0

The *Journal of Transport and Land Use* is the official journal of the World Society for Transport and Land Use (WSTLUR) and is published and sponsored by the University of Minnesota Center for Transportation Studies.

the total trip ends, rather than changing them arbitrarily.

The second approach also induces a distortion of real data as the change of the deterrence function in the trip distribution model leads to changes in the passenger trip length distribution. However, the current level of knowledge about the TLD causes most researchers to believe that the distortion is quite justifiable as a sampling survey provides some evaluation of the function only. The real variant of the TLD function can considerably differ from the function defined by the survey results. Unfortunately, the sample theory does not provide transport engineers with recommendations to estimate acceptable variations of the number of trips within the bins of trip length or time distribution. It forces a researcher to collect a considerable amount of information about the distance and time of trips that brings about large costs but fails to provide the guaranteed validity of the TLD estimation.

It is possible to improve the tools of demand modeling by means of the determination of a universally applicable function that describes the TLD. Then the task of a sample survey will be to determine only the moments of the TLD. In this case, the sample theory provides a reliable tool set. Such a way to determine the TLD will allow us to simultaneously decrease the costs of conducting the survey and get a reliable basis to distribute trips in the O-D matrix. The use of the TLD function as a constraint on trip distribution will enable us to significantly decrease the ambiguity of the O-D matrix since the TLD is a stable statistic regularity which reflects real transport demand.

To identify a model that is suitable for the description of the TLD in most cities, it is expedient to determine the prerequisites of the nature of regularities in travel distances in the cities. To solve the task, it is reasonable to use the characteristics of the transport infrastructure as the information basis for the research.

2 Literature review

A lot of researches are devoted to population mobility in cities. In many cases the focus of the study centered on home-based work (HBW) trips as they are a significant part of daily activity (Ahern et al., 2013). This type of travel obviously has the most significant impact on the choice of the place of residence and work. Some researchers and research organizations have obtained the theoretical models that can be used to describe the TLD. In most cases, they encountered the problem in choosing a function that would describe an empirical distribution similar to the gamma (or Weibull) distribution. Thus, the authors (Ortuzar & Willumsen, 2011) pointed to such a distribution as typical for urban population (see Figure 1).

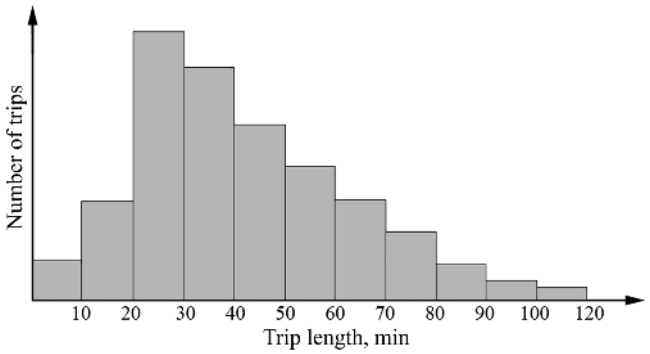


Figure 1: Typical plot of trip length distribution. Adapted from Modelling Transport (p. 184), by J.D. Ortuzar and L.G. Willumsen, 2011, Chichester: John Wiley & Sons. Copyright 2011 by the John Wiley & Sons.

The typicality of the graph of the TLD function is confirmed in the works of Benson, Teniente, Stover, and Cunagin (1979), Bovy, Bliemer, and van Nes (2006), Moeckel (2017), Zhao, Chow, Li, and Gan (2004) as well as by the results of the empirical research on the TLD function that relates to various time periods, separate businesses, cities and countries (Aultman-Hall, Sears, Dowds, & Hines, 2012; Chacaga, Rudnicki, & Sroka, 2010; Englund, Eash, & Lupa, 2010; Katsis, Papageorgiou, & Ntziachristos, 2014; Veenstra, Thomas, & Tutert, 2010). For example, a typical TLD plot was obtained by processing the data that had been collected by Air Sage in The Research Triangle in North Carolina, the USA (Huntsinger & Donnelly, 2014). Similar TLD curves were obtained in Alexandria, Egypt (Mounir, 2014), Lincoln, Nebraska, the USA (L & A Transportation [LAT], 2006), and Seoul, South Korea (Kim & Lee, 2001). The graphs of the TLD function, which were obtained in Moscow and Saint Petersburg, Russia (Efremov & Golc, 1988; Shelejhovskij, 1946), and in Prague and Plzen, Czech Republic (Cibulka, 1987) by researchers of the former Eastern bloc countries, externally look like the gamma distribution.

The shape of the TLD curves is obviously similar although their description is made by different functions: the exponential or power function with negative exponent (Yang, Jin, Wan, Li, & Ran, 2013), the Erlang distribution function (Cibulka, 1987), the log-normal distribution function (Katsis, Papageorgiou, & Ntziachristos, 2014), the gamma distribution function (Benson et al., 1979; Transportation Research Board [TRB], 2010; Yang et al., 2013), etc. It indicates that there is no theoretical validation of a general shape of the TLD function even if empirical data are available.

Most papers, which deal with the TLD, contain the description of the use of the distribution to calibrate the estimated trips in the O-D matrix (Aultman-Hall et al., 2012; Englund et al., 2010; Fricker & Jin, 2008; Huntsinger & Donnelly, 2014; Kim & Lee, 2001; LAT, 2006; Mounir, 2014; Veenstra, Thomas, & Tutert, 2010; Yang et al., 2013) and few papers are devoted to research on the factors that influence the distribution. They indicate a connection between the characteristics of land use in a city and the length or frequency of trips (Acheampong & Silva, 2015; Benson et al., 1979; Gehrke & Clifton, 2016; Junge & Levinson, 2012; Milakis, Cervero, & Wee, 2015; Porter, Brown, Dunphy, & Vimmerstedt, 2013; Srinivasan, Provost, & Steiner, 2013; Zhao et al., 2004). The work by Benson et al. (1979) states that the time of trips increases if city population grows, and it decreases if the travel speed and the concentration of business go up at downtown. Researchers Stead and Marshall (2001), Wegener and Fuerst (2004) state that the farther a place of residence from the city center is, the longer HBW trips are. The work (Yigitcanlar, Dodson, Gleeson, & Sipe, 2005) reads that if the autonomy of separate districts in the city grows, the travel distance of population goes down.

The papers mentioned indicate that the TLD is the function of land use and city planning, but the explanation of the reasons for a general view of the TLD plot is absent. At the same time, each empirical TLD was obtained using the function that was deemed the most appropriate for its description. This process is not supported by any theoretical prerequisites and comes as a simple choice of the theoretical distribution function on the base of goodness-of-fit tests.

The above-mentioned information points to the fact that the functions are chosen on the basis of researchers' considerations only. They cannot serve as a substantial theoretical validation of current regularities of HBW trip lengths. The research results, which are presented in the papers reviewed, rather persuasively indicate intrinsic regularities in urban transport systems that bring about unambiguity in the distribution of HBW trip length. These regularities are little-studied and they are of great interest for researchers to clarify the reasons for TLDs similarity.

Theoretical validation of the TLD shape and determination of a universal mathematical function for HBW trip length distribution on the basis of a mass transit infrastructure in the city enable us

- to contribute to the development of methods to get actual TLD functions, as it will be sufficient just to identify the parameters of trip length distribution. It will enable to significantly decrease

requirements to the data that are collected during transport surveys;
- to check the trip distribution models that are mainly based on the travel cost while the probability of a trip, which is determined by the TLD function, has to be reproduced by the O-D matrix;
- to provide the basis to create new trip distribution models for transport demand modeling.

3 Fundamentals

3.1 Trip length distribution as outcome of the distance matrix and the O-D matrix

HBW trips are studied well and they are the most stable segment of the urban passenger transportation market (Ortuzar & Willumsen, 2011). In addition, if compared with other trips, they are considered to have the most significant impact upon the choice of the place of residence (Moeckel, 2017; Xie & Levinson, 2011).

The aim of the paper is to investigate the reasons of population settlement regularities and estimate the role of the O-D matrix in the formation of the TLD function. Therefore, the object of the study is the distribution function of HBW trip length.

This paper assumes that the basis to determine the settlement regularities reflected by the HBW TLD is the hypothesis about the random location of origins and destinations (trip attractors) when the city population and area increase under the common people aspiration to live and work closer to the city center. On the one hand, the hypothesis is based on a large number of factors that influence the choice of work and a residence location by person (Milakis et al., 2015). It is important to emphasize that a variety of factors results in the comparatively low influence of a separate factor on the full set of "residence-work"pairs in the city. In the end, it results in the stable regularities of HBW trip lengths (Shelejhovskij, 1946) that can be described by theoretical density functions. On the other hand, the hypothesis corresponds to a well-known tendency for real estate prices: the greater the distance from the city center, the lower the price (Xie & Levinson, 2011). This tendency is partly determined by the increase in the opportunities to achieve a travel purpose in the city center (Horner & Downs, 2014).

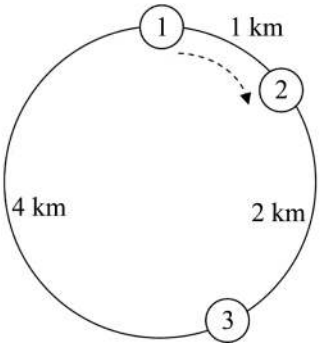
The objectives of this paper require the discretization of the city area to identify origins and destinations. There are two alternatives in the city transport model: public transit (PT) stops and transportation zones. The transportation zones are universal though they reflect the subjective views of city planners. From this point of view, the location of stops in the city is strictly determined and completely objective. It allows us to avoid any subjectivity when developing the theoretical fundamentals of the formation of the TLD function that is really important for the study of random processes. In this case, the stop locations can be interpreted as the places of demand generation and attraction. Therefore, the information basis to develop the theoretical fundamentals of the TLD function formation and experimentally verify them is grounded on the spatial characteristics of PT stops. So, this paper is confined to the investigation of the TLD for PT passengers only, excluding any options to travel by private transport or on foot. The use of the stop coordinates to discretize the city area limits the application of the research results to the in-vehicle component of the HBW trip. Any other components are out of the analysis. Consequently, the theoretical background will be developed for the in-vehicle TLD, which is an important distribution for the PT demand modeling.

The source of the empirical data to develop the TLD function in this paper is the model of a PT network. We assume that the mass transit infrastructure corresponds to the transport demand of the city population. So, a passenger's trip distance can be regarded as the shortest route between origin and destination stops and the parameters of the TLD function are assumed to be independent of the route network. This condition is a priori as current changes of routes can hardly bring about tangible migra-

tion processes in the city. In addition, both mentioned assumptions are aimed at obtaining an objective characteristic of the spatial remoteness of trip origins and destinations—it is the shortest in-vehicle distance between the origin and destination stops.

A passenger trip length per se is a very specific random variable. At first, it seems to be continuous; however, it is actually discrete. The TLD function reflects the probability of the finite set of the distances between origins and destinations. The number of distinct distances depends on the number of cells in the matrix of distances between stops and is limited by the value $(N^2 - N)$, where N is the quantity of transit stops in the city (i.e., the number of rows and columns in the matrix of distances). The N distance values on the diagonal are subtracted from N^2 because intra-zonal trips are ignored in this paper. A real quantity of distinct distances can be less than $(N^2 - N)$ because the matrix of the shortest distances can contain equal values. However, the discreteness of the TLD does not negate the possibility to describe it as a continuous random variable, given a sufficiently large number of values.

Now we discuss the mechanism of the transformation of one random variable—the shortest distance between stops—into another random variable—the distance of HBW trips. In order to do that, we can slightly simplify the situation with a suggestion that there are no equal values among distances between various pairs of stops (origins and destinations). Then the probability of each distance to appear in a total set of trip lengths consisting of H values is equal to $P(l_{ij}) = h_{ij}/H$, where l_{ij} is the distance between stops i and j, h_{ij} is the number of trips between stops i and j, and H is a total number of trips in the O-D matrix. It means that the number of trips between stops i and j directly determines the probability of the distance l_{ij} between i and j in the set of HBW trip distances. It can be clearly seen from the information presented in Figure 2. Let a simulated object consist of three stops that are located at a ring road with one-way clockwise traffic. The distances between the stops are in the distance matrix. During the period under review twelve trips are made according to the O-D matrix adduced. The histograms of the distribution of the distances between stops and passenger trip lengths, which correspond to the matrices, are shown in Figure 3.



Distance matrix, km

Origins	Destinations		
	1	2	3
1	0	1	3
2	6	0	2
3	4	5	0

O-D matrix

Origins	Destinations			
	1	2	3	∑
1	0	1	3	4
2	1	0	2	3
3	3	2	0	5
∑	4	3	5	12

Figure 2: Initial data to illustrate the transformation of the distances between stops into TLD

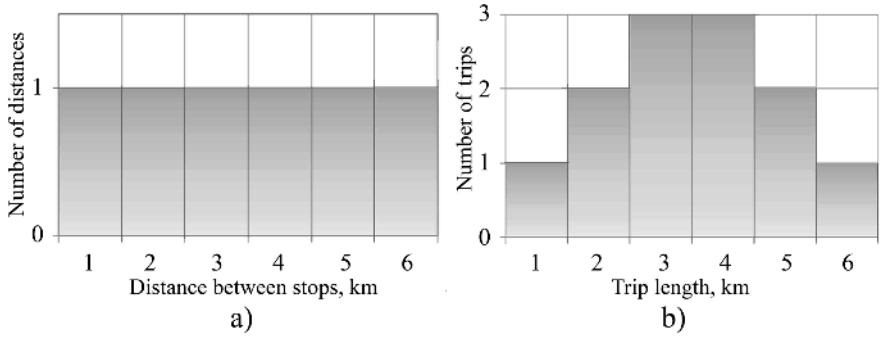


Figure 3: Distribution of: a) distances between stops from the matrix of distances; b) trip lengths

In this example, the distances between stops correspond to the rectangular distribution (see Figure 3a), and the passenger trip lengths correspond to the triangular distribution (see Figure 3b). So the final distribution of passenger trip lengths depends on the distances between stops and their further transformation using the O-D matrix. The essence of the transformation is that the distances between stops are transformed into trip distances by means of the non-negative number of repetitions of each distance.

Looking into the causes that bring about the well-known TLD function, it is first necessary to find out the distribution of the distances between stops in real cities and then estimate the results of the transformation of the distances using the O-D matrices generated by various trip distribution models.

3.2 Research of distances between pairs of stops as the sum of link lengths between adjacent stops

The route between any pair of PT stops is a set of links between adjacent stops (hereinafter referred to as links). This paper assumes that TLD function for PT passengers is based on the shortest distances between stops which, in turn, are determined as the sum of the link lengths

$$l_{ij} = \sum_{k=1}^{n_{ij}} l_k \tag{1}$$

where:

l_{ij} Distance between stops i and j
n_{ij} Number of links between stops i and j
l_k Length of link k using the shortest route from i to j

In this case, the walking distance to and from the stop is neglected that can be regarded as a consequence of the use of stop coordinates for the discretization of the city area. It is partly justified because the access distance is usually much shorter than the distance traveled in a vehicle. According to the current standard (TRB, 2013), stops should be within the access distance, which is rather short. Some research indicates the shortening of the walking distance as transit trips become shorter (Daniels & Mulley, 2013; El-Geneidy, Grimsrud, Wasfi, Tétreault, & Surprenant-Legault, 2014; O'Sullivan & Morrall, 1996). Thus, the disregard for the walking distance contributes to the limitation of the research results by the in-vehicle component of the HBW trip.

The number of links along the way between stops, as well as the length of each link, is the result of the long-term evolution of the city territory and the parallel development of its infrastructure, including the PT network. Any decision to set up a new stop is, on the one hand, subjective and, on the other hand, is based on an economic, planning and transportation situation at the moment of decision-making. Therefore, the results of the decisions can be considered random for people who have not taken part in decision-making. Accordingly, in this context the coordinates of each stop can be considered random. In consequence, the number of links along the shortest route between stops and the length of each separate link can also be considered as random variables. Thus, random values of the shortest distances between stops can be considered as the sum of the random number of random components.

To determine the characteristics of the length and the number of the links along the route between stops, the process of the city territory expansion should be considered. It is expedient to do it in terms of our hypothesis about the random deployment of origins and destinations during the expansion of the city territory. In terms of the probability theory the hypothesis can be formalized as the symmetrical scattering of the horizontal coordinates of stops around the city center. This rule will only work in approximately equal geographic and administrative conditions in the territories around the built-up area of the city.

For a "typical" city it can be logical to suppose that the scattering of the stop coordinates will be approximately normal. In this case the distribution of the straight-line distances between the PT stops and the city center will correspond to the Rayleigh distribution

$$F(d_j) = 1 - e^{-\frac{d_j^2}{2\sigma_d^2}} \tag{2}$$

where:

d_j Straight-line distance between the stop j and the city center

σ_d Rayleigh distribution parameter concerning the variable d_j (Chow & Teiher, 1978; Feller, 1966)

The city center should be a point in the historical center area from where the city area expansion began. In particular, it can coincide with one of the PT stops. The hypothesis will be considered consistent with reality if the empirical distribution of d_j corresponds to the Rayleigh distribution.

Provided that the hypothesis is confirmed, the link length can be regarded as a variable that has been determined by regularly repeated events when the transport system develops and new transit stops appear. In this case, a link can be presented as the distance between a stop which is beyond the circumference drawn around all the existing normally distributed stops and a stop on the circumference (see Figure 4).

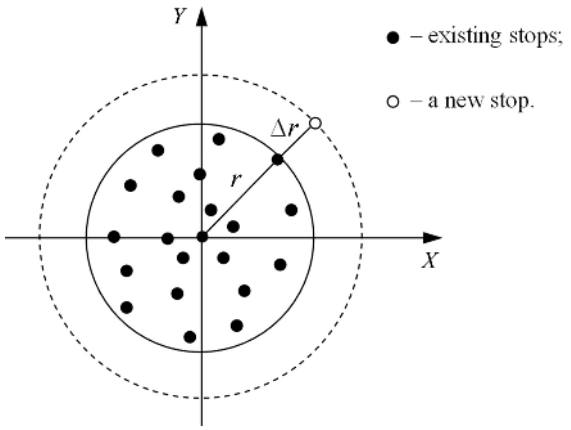


Figure 4: Graphic presentation of a new link appearance

It should be noted that value Δr is the width of a certain ring and can be regarded as a minimal length of the link between an existing stop and a new one, which has been set up as a result of the development of the city transport infrastructure. This length appears when a new stop is located on the same radius as an existing stop (see Figure 4). In other cases, the length of the link will be rather longer.

To study the properties of a minimal link length, it is expedient to introduce the random variable $R = (r+\Delta r)$, which is a straight-line distance from a new stop to the center of the scatter (city center). The distribution function of the given random variable is equal to a probability that its value will exceed the radius r of the circumference drawn around the existing stops. As the probability of a new stop appearing within the mentioned circumference is determined by the Rayleigh distribution function, the distribution function of the variable R is as follows

$$F_R = P(E) = P\{R > r\} = 1 - P\{R < r\} = 1 - (1 - e^{-\frac{r^2}{2\sigma_r^2}}) = e^{-\frac{r^2}{2\sigma_r^2}} \tag{3}$$

where:

E Event that a new stop will appear outside the circumference with radius r

σ_r Rayleigh distribution parameter concerning the variable R

Then it is possible to find a probability that a new stop will be outside the circumference of radius r but within the circumfrence of radius $(r + \Delta r)$. The probability indicates that a stop will appear within the ring between smaller radius r and bigger radius $(r + \Delta r)$:

$$P(J) = P\{(R > r) \cap (R \le r + \Delta r)\} = P\{R > r\} - P\{R > r + \Delta r\} = \\ = e^{-\frac{r^2}{2\sigma_r^2}} - e^{-\frac{(r+\Delta r)^2}{2\sigma_r^2}} = e^{-\frac{r^2}{2\sigma_r^2}}(1 - e^{-\frac{r\cdot\Delta r}{\sigma_r^2} - \frac{\Delta r^2}{2\sigma_r^2}}) = P(E)\cdot(1 - e^{-\frac{r\cdot\Delta r}{\sigma_r^2} - \frac{\Delta r^2}{2\sigma_r^2}}) \quad (4)$$

where:

J Event that a new stop will be within the ring

The process of the transportation network development implies that a new stop will appear outside the inner radius of the ring. Therefore, provided event E is a persistent event $(P(E)=1)$, the probability of event J can be given as follows

$$P(J|E) = 1 - e^{-\frac{r\cdot\Delta r}{\sigma_r^2} - \frac{\Delta r^2}{2\sigma_r^2}} = 1 - e^{-\frac{2r\cdot\Delta r + \Delta r^2}{2\sigma_r^2}}. \quad (5)$$

Because of the independence and identical normal distribution of the stop coordinates, the conditional probability presented by Equation 5 coincides with a conditional probability that a stop will be located within a certain segment of the ring at any radius under the condition that the stop will be outside the inner circumference of radius r. It stipulates a transition from a squared radius to its linear form in Equation 5.

For further study of parameter Δr, i.e., a minimal link length, it is necessary to know its distribution. It should be noted that $\Delta r \approx r$ is true in the city center only, i.e., at the very beginning of city growth. Under those conditions, movements are mainly made on foot. For the mass transit, it is reasonable to assume that $\Delta r << r$. Then it can be accepted that $\Delta r^2 \approx 0$and we can rewrite Equation 5 as

$$P(J|E) = F(l_{k\min}) = 1 - e^{-\frac{2r}{\sigma^2}\Delta r} = 1 - e^{-\lambda_r\cdot\Delta r},\ \lambda_r = 2r/\sigma_r^2 \quad (6)$$

where:

$l_{k\min}$ Minimal link length, i.e., the length of a direct segment between two points (stops) on the same radius that starts at the city center

The right part of the equation is a cumulative function of the exponential distribution with parameter $\lambda = 2r/\sigma_r^2$. As the conditional probability $P(J|E)$ is the distribution of a minimal link length, it follows from Equation 6 that it must be exponential. So the conformity between the Rayleigh distribution and distances d_j between the city center and each PT stop brings about a potential possibility to describe a minimal link length $l_{k\min}$ by the exponential distribution.

It should be noted that a minimal link length is determined not only by the straightness of the line, which links two stops. To a considerable extent, it is formed by a mutual position of adjacent stops, but it contradicts the hypothesis about the randomness of the PT stops scattering during the city growth. Therefore, the conclusion about the exponential distribution of minimal link lengths cannot be final either for a link length or for the opportunity to assume that the location of a stop is random. Moreover,

a single case of non-refutation of the hypothesis cannot be considered sufficient to prove it. Therefore, the hypothesis about an exponential distribution of link length needs additional verification.

For the in-depth study of the stop location it is worth to discuss a mutual location of adjacent stops, i.e., the link endpoints. The question to be answered is about the location of the link ends if all the link starting points are placed at one point. If a stop's location is random, the relative coordinates of the link ends will have the distribution that is close to normal. To check the statement, the Rayleigh distribution (Equation 2) can help again, but the other parameter is necessary, i.e., the variance of link length σ:

$$F(l_k) = 1 - e^{-\frac{l_k^2}{2\sigma^2}}. \tag{7}$$

The correspondence between the link lengths and the distribution will additionally confirm the hypothesis about the random stop locations in the city. Prima facie, such confirmation contradicts the preliminary conclusion about the exponential distribution of $l_{k\,\min}$. However, the contradiction is settled by the property of the Rayleigh distribution that it corresponds to the squared exponential variable. Therefore, it is necessary to move from a squared link length to a linear one. It can be done by means of linearization using expansion in a Taylor series (Saff & Snider, 1976), see Appendix A:

$$F(l_{k\,\text{adj}}) = 1 - e^{-\lambda(l_k - \Delta l)} = 1 - e^{-\lambda l_{k\,\text{adj}}} \tag{8}$$

$$l_{k\,\text{adj}} = l_k - \Delta l \geq 0 \tag{9}$$

where:

$l_{k\,\text{adj}}$ Random component of link k, km

λ Distribution parameter that equals $1/2\sigma 2$

Δl Shift parameter of factual link lengths l_k, km

Shift parameter Δl can be taken equal to minimal link length $l_{\min}$ in the city. It is explained in Appendix A. In this case, a real link length having a shift parameter is presented as

$$l_k = l_{k\,\text{adj}} + l_{\min} \tag{10}$$

where:

$l_{\min}$ Both the minimal link length in the city and the shift parameter of the random variable $l_{k\,\text{adj}}$, km

Taking into account Equation 1, the distance between any pair of stops in the city can be written as

$$l_{ij} = \sum_{k=1}^{n_{ij}} l_k = \sum_{k=1}^{n_{ij}} (l_{k\,\text{adj}} + l_{\min}) = \sum_{k=1}^{n_{ij}} l_{k\,\text{adj}} + \sum_{k=1}^{n_{ij}} l_{\min} = \sum_{k=1}^{n_{ij}} l_{k\,\text{adj}} + n_{ij} \cdot l_{\min}. \tag{11}$$

This equation is essential for the study of the distribution characteristics of the distances between stops. So, it is necessary to thoroughly investigate theoretical prerequisites of the distribution law of the variable $l_{k\,\text{adj}}$. The transformation of a Rayleigh distributed random variable into an exponential variable just illustrates the relationship between two theoretical distributions. At the same time, the successive appearance of new single stops during the city growth idealizes a real process of the city evolution. The expansion of the city is usually accompanied with the emergence of several stops rather than one stop.

To determine the theoretical prerequisites of the $l_{k\,\mathrm{adj}}$ distribution, a more realistic variant of link formation should be considered. The essence of the variant is in the assignment a number of intermediate stops between an existing final stop in the PT network and a new final stop on a new territory. Mathematically the variant is provided in Appendix B when using two approaches in order to research the sum of $\sum_{k=1}^{n_{ij}} l_{k\,\mathrm{adj}}$ from Equation 11. This sum is denoted as $l_{ij\,\mathrm{adj}}$, i.e., $l_{ij\,\mathrm{adj}} = \sum_{k=1}^{n_{ij}} l_{k\,\mathrm{adj}}$.

Thus, the hypothesis of the exponential distribution of $l_{k\,\mathrm{adj}}$ for the process of the assignment of several intermediate stops between an existing final stop and a new final stop on a new territory in the PT network makes sense. This fact is a basis for further study of the mechanism of TLD formation.

3.3 Theoretical validation of link length distribution

The in-depth analysis of Equation 11, which is used to represent a distance between a pair of stops *i* and *j*, enables us to state that, when determining the distribution parameters of l_{ij}, the use of sum $l_{ij\,\mathrm{adj}}$ itself can cause inaccuracies. The reason for possible inaccuracies is an unequal frequency of the "inclusions"of certain links into the shortest routes between the pairs of stops. If link lengths have a rectangular distribution, this problem is not available. However, for an exponential variable, the probability of its occurrence in a sample decreases as a link length increases. The number of the "inclusions" of long links into the shortest routes between stops will probably decrease because the routes are formed as the samples from an existing set of links having a certain distribution. Therefore, it is necessary to theoretically determine the characteristics of the distribution of link lengths when considering links as the components of the routes between various pairs of PT stops. These routes are formed as the samples of different size from the population of links having the known distribution.

In this case, the primary hypothesis is the assumption that, when forming the samples from the population of links, the probability of the occurrence of a certain value in a sample will depend on its frequency in the population: the higher the frequency, the higher the expected probability. The converse of the assumption is also true: low-frequency links in the population will be less frequent in the samples.

It is expedient to study this issue in terms of the probability theory. Corresponding mathematical transformations are provided in Appendix C, where a new random variable $l_{s\,\mathrm{adj}}$—the length of *s*-th link along the route between stops *i* and *j*—is considered. According to analytic calculations in the Appendix, the following statement is true: if the link length distribution is exponential, then short links will appear more frequently than long ones on the routes between all the pairs of stops.

This fact lets us conclude that the type of the distribution of link lengths, which constitute l_{ij}, does not differ from the type of the distribution of $l_{k\,\mathrm{adj}}$, but have the different slope of the density function curve. The slope can be characterized by a greater value of the exponential distribution parameter $\lambda_s > \lambda$, where λ_s is the parameter of the predicted exponential distribution of the values $l_{s\,\mathrm{adj}}$, and λ is the parameter of the initial exponential distribution of $l_{k\,\mathrm{adj}}$.

At the same time, in addition to the distribution type, the number of values in initial and transformed sets of links changes as well, i.e., during the formation of the set of routes between stops the transformation from the random variable $l_{k\,\mathrm{adj}}$ into the random variable $l_{s\,\mathrm{adj}}$ takes place. Therefore, it is necessary to experimentally check if two sets of links follow an exponential distribution as well as compare distribution parameters. The decrease of an average link length in the set of $l_{s\,\mathrm{adj}}$ relative to the set of $l_{s\,\mathrm{adj}}$ will confirm the inferences above.

3.4 Determining the distribution of the distances between pairs of stops

The next undefined variable in Equation 11 is the number of links along the shortest routes between the

pairs of stops *i* and *j*. As the variable is formed under the influence of a number of factors (regulatory requirements on link lengths, city layout, ability and expediency of transit stop construction subject to existing transport demand etc.), it can hypothetically have a near-normal distribution. Since the upper limit of a sum can have discrete values only, the number of the links along the shortest route between stops must be a discrete variable. The closest to a normal distribution are Poisson and triangular discrete distributions that are widespread in engineering applications of the probability theory. We describe the distribution of the number of links by two mentioned distributions.

Considering the number of links along the route of random length, the Poisson distribution is as follows

$$p(n_{ij}) = \frac{\mu^{n_{ij}}}{n_{ij}!} e^{-\mu} \quad (n_{ij} = 0, 1, 2, ...) \tag{13}$$

where:

μ Poisson distribution parameter

The Poisson distribution becomes symmetrical and can be approximated by the normal distribution only under the large values of parameter μ. The parameter of the Poisson distribution is the mean that is equal to the variance $\mu=\sigma^2$ (Forbes, Evans, Hastings, & Peacock, 2011). Under real values of the number of links along the routes around the city, the symmetrical Poisson distribution will have a large variance. To make it possible to describe the real number of links by the Poisson distribution, it is necessary to introduce a shift parameter $n_0=const$ into the distribution. It will enable us to get a symmetrical distribution and obtain the correspondence between mean and variance, which is a property of the Poisson distribution. The correspondence between the Poisson distribution and the distribution of n_{ij} needs the estimation when each value n_{ij} is increased by a certain constant to get the large values of μ without the change of variance. This method of getting information about the regularity in the values of n_{ij} should be validated using empirical data.

The triangular distribution has a series which in the case of a maximal odd value of n_{ij} is represented by the expression

$$p(n_{ij}) = \begin{cases} 0, \ n_{ij} \le 0, \ n_{ij} > 2N_c + 1; \\ \dfrac{n_{ij}}{(N_c+1)^2}, \ 1 \le n_{ij} \le N_c + 1; \\ \dfrac{2N_c - n_{ij} + 2}{(N_c+1)^2}, \ N_c + 1 < n_{ij} \le 2N_c + 1 \end{cases} \tag{14}$$

where:

N_c Parameter of the triangular distribution

In the case of a maximal even value of n_{ij}, the triangular distribution is as follows:

$$p(n_{ij}) = \begin{cases} 0, \ n_{ij} \le 0, \ n_{ij} > 2N_c; \\ \dfrac{n_{ij}}{N_c(N_c+1)}, \ 1 \le n_{ij} \le N_c; \\ \dfrac{2N_c - n_{ij} + 1}{(N_c+1)^2}, \ N_c < n_{ij} \le 2N_c. \end{cases} \tag{15}$$

The parameters of the distribution of distances between stops, which are obtained using the Poisson

and triangular distributions, can be regarded as rough estimates rather than a final evaluation result because of the non-complete theoretical validation of the application of distributions presented by Equation 13, Equation 14, and Equation 15.

Equation 11 consists of two parts, and each part has the variable n_{ij}. The first part of Equation 11—$\sum_{k=1}^{n_{ij}} l_{k\,\mathrm{adj}}$ —will be the convolution of the exponential distribution functions of link lengths $l_{k\,\mathrm{adj}}$. The result of such a convolution is the Erlang distribution of n_{ij} degree (Feller, 1966; Forbes et al., 2011). So the distribution function of $l_{ij\,\mathrm{adj}}$ under the values greater than 0 is the probability mixture of the Erlang distributions of 1-st, 2-nd, ..., n_{ij}-th degree with probabilities p_1, p_2,.....p_n. The second component of Equation 11—$n_{ij} \cdot l_{\min}$—will have the same distribution as variable n_{ij} because $l_{\min}$=*const* for each city.

The density function of the variable l_{ij} can be diverse and depends on the distribution applied to describe the variable n_{ij}. Analytical transformations that allow obtaining the density function of the variable l_{ij} are possible for all variants of the distribution of the quantity of links that are expressed by Equation 13, Equation 14, and Equation 15. These analytical transformations are not presented within the paper.

When describing the variable n_{ij} by the Poisson distribution, the density function of l_{ij} will be as follows:

$$f(l_{ij}) = \sum_{n_{ij}=1}^{\left[\frac{l_{ij}}{l_{\min}}\right]} f_{n_{ij}}(l_{ij}) \cdot \frac{\mu^{n_{ij}}}{n_{ij}!} e^{-\mu} (1 - e^{-\mu})^{-1} \tag{17}$$

where:

$$f_{n_{ij}}(l_{ij}) = \begin{cases} \dfrac{(\lambda_s)^{n_{ij}} (l_{ij} - n_{ij} \cdot l_{\min})^{n_{ij}-1}}{(n_{ij}-1)!} e^{-\lambda_s (l_{ij} - n_{ij} \cdot l_{\min})}; \; l_{ij} \geq n_{ij} \cdot l_{\min}; \\ 0; \; l_{ij} < n_{ij} \cdot l_{\min}. \end{cases}$$

In the case when the variable n_{ij} can be described by the triangular distribution having even maximal value, the density function of l_{ij} will be as follows:

$$f(l_{ij}) = \frac{1}{N_c^2} \sum_{h=1}^{\min\left[\frac{l_{ij}-l_{\min}}{l_{\min}},\, N_c\right]} \sum_{u=1}^{\min\left[\frac{l_{ij}-h \cdot l_{\min}}{l_{\min}},\, N_c\right]} \frac{(\lambda_s)^{h+u}}{(h-1)!} e^{-\lambda_s (l_{ij} - (h+u) \cdot l_{\min})} \left(l_{ij} - (h+u) l_{\min}\right)^{h+u-1} \times \\ \times \sum_{d=0}^{u-1} (-1)^d \cdot \frac{1}{d!(u-1-d)!(h+d)}. \tag{18}$$

If n_{ij} is described by the triangular distribution having an odd maximal value, the density function of l_{ij} will be as follows:

$$f(l_{ij}) = \frac{1}{N_c(N_c+1)} \sum_{h=1}^{\min\left[\frac{l_{ij}-l_{\min}}{l_{\min}},\, N_c\right]} \sum_{u=1}^{\min\left[\frac{l_{ij}-h \cdot l_{\min}}{l_{\min}},\, N_c+1\right]} \frac{(\lambda_s)^{h+u}}{(h-1)!} e^{-\lambda_s (l_{ij} - (h+u) \cdot l_{\min})} \times \\ \times \left(l_{ij} - (h+u) l_{\min}\right)^{h+u-1} \sum_{d=0}^{u-1} (-1)^d \cdot \frac{1}{d!(u-1-d)!(h+d)}. \tag{19}$$

The theoretical study of the components of Equation 11 has resulted in the variants of the density function of the distances between stops l_{ij}. The obtained density functions of l_{ij} definitely do not coincide with any known distribution law. At the same time, the functions are not completely stipulated by the urban transport infrastructure as they are based on the distributions of n_{ij} that have no reliable theoretical validation. They are a certain approximation of the possible distribution of distances between stops. They can be substituted if the most appropriate theoretical distribution function is found among

well-known two-parameter distributions. The necessity for two parameters in the required distribution is conditioned with two parameters λ_s and μ (or N_c) in the obtained variants of $f(l_{ij})$. The search among the most widespread theoretical distributions for one that can approximate or substitute the density functions presented by Equation 17, Equation 18, and Equation 19 is performed graphically. It is done by means of the curves of the density functions having various values of parameters λ_s, μ, and N_c. These curves are in Figure 5, Figure 6, and Figure 7.

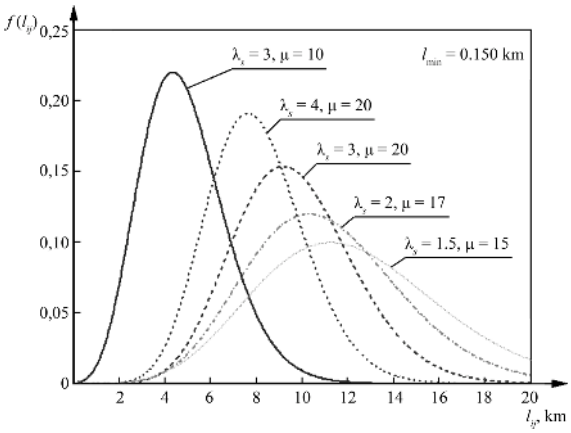


Figure 5: Density function of the distances between all possible pairs of transit stops when the number of links is Poisson distributed

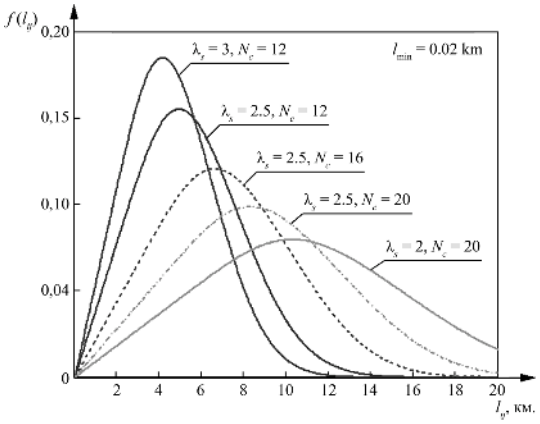


Figure 6: Density function of the distances between all possible pairs of transit stops when the number of links is triangular distributed and the maximal value of the number of links is even

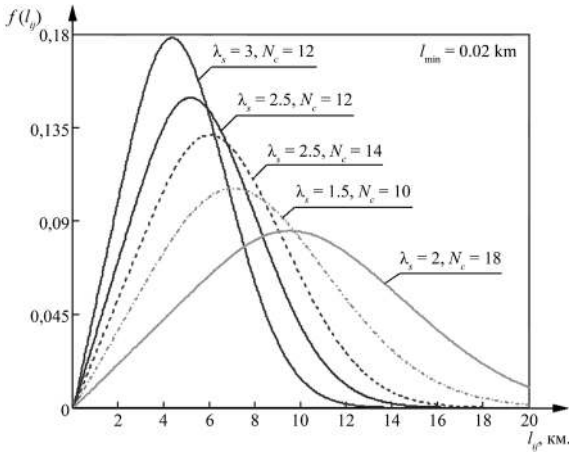


Figure 7: Density function of the distances between all possible pairs of transit stops when the number of links is triangular distributed and the maximal value of the number of links is odd

The obtained curves are very similar to the gamma and Weibull distribution curves. Taking into account the fact that the distributions presented in Figure 5, Figure 6, and Figure 7 are probabilistic mixture of Erlang distributions and that the Erlang distribution is a special case of the gamma distribution, the next hypothesis can be put forward: if the assumed distribution of variable n_{ij} is confirmed, then the variable l_{ij} will correspond to the gamma distribution.

If this hypothesis is confirmed, it will testify that TLD regularities are stipulated by stop locations in the city rather than by the O-D matrix. In other words, the obtained theoretical results testify that a typi-

cal TLD plot is not the result of the trip distribution in the matrix. It is formed by the spatial location of trip attractors in the city. This situation refutes, to some extent, a common approach to define trip patterns that is based on a considerable impact of transportation factors on the travel demand formation.

The above-mentioned statements should be verified experimentally. Experiments will also clarify the issue of the influence of the O-D matrix upon the transformation from the distribution of distances between stops to the trip length distribution.

4 Experimental research

4.1 Verification of the hypothesis about the exponential distribution of link lengths

The basis for the experimental evaluation of the theoretical fundamentals for the formation of the TLD is the transport models of the Ukrainian cities of Kyiv, Kharkiv, Sumy, Kryvyi Rih (Dnipropetrovsk region), Kirovohrad, Sverdlovsk (Luhansk region), Oleksandrija (Kirovohrad region), Izium, Balaklia and Kupiansk (Kharkiv region). The models are developed in the framework of projects and research in the Department of Transport Systems and Logistics in Kharkiv National Automobile and Highway University, Ukraine, using the VISUM software of the German company PTV AG (VISUM, 2010).

These cities represent almost all groups of classification by the criterion of population: Kyiv and Kharkiv have a population of over 1 million; Kryvyi Rih—just over 650,000; Sumy—270,000; Kirovohrad—just over 230,000; Sverdlovsk, Alexandrija and Izium belong to the group of cities with a population between 50,000 and 100,000; Balaklia and Kupjansk have a population under 50,000. The detailed characteristics of the cities, including the area and transit network topology, are presented in Appendix D. The transport models of the cities allow us to get accurate data: stop coordinates, link lengths and the matrices of distances between stops.[1] The use of the data will enable us to prove the universality of the developed theoretical fundamentals and their applicability to most cities having the PT.

In order to avoid a number of identical graphs presenting the empirical distribution and the fitted curve, this section will give an example of such a graph for one city at each stage of experimental research. The curves for the rest of cities will be in two different graphs: one will illustrate the curves for the cities with a population over 250,000 and another—less than 250,000. The cities are grouped to reach similarity and approximately equal number of curves in the figures. Theoretical distribution parameters and goodness-of-fit measures will be in one table for all the cities. To evaluate the goodness of fit, the χ^2-test is chosen as one of the most common tests used during distribution fitting (Chow & Teiher, 1978; Feller, 1966; Forbes et al., 2011). The distribution parameters are given in order to provide an opportunity to estimate the frequency moments, which are the main characteristics of the route networks of the cities (Harznagy, Fi, London, & Nermeth, 2015). At the same time, all the parameters are to approve of the theoretical findings in Fundamentals—they point to the fact that in spite of certain differences between parameter values within separate items each theoretical issue is confirmed in a wide range of cities.

The first step is to check the compliance between the Rayleigh distribution and straight-line distances from a city center ("central" stop) to all other stops. All stops in the cities were displayed in the map in VISUM. Then a circumference was drawn around most stops. A "central" stop is the closest to the crossing of the orthogonal diameters of the circumference. Such a "central" stop in the cities is located close to the historical city center. The target distances were calculated as follows:

$$d_j = \sqrt{(X_j - X_{cent})^2 + (Y_j - Y_{cent})^2}, \tag{23}$$

where:

[1] All data used during Experimental Research are available online at The Characteristics of Public Transit Networks and Population Trip Lengths in Ukrainian Cities, 2008-2016 (http://doi.org/10.3886/E100393V1), by P. Horbachov and S. Svichynskyi, 2017.

X_j, X_{cent} Abscissas of stop *j* and a "central" stop correspondingly
Y_j, X_{cent} Ordinates of stop *j* and a "central" stop correspondingly

The parameter of the Rayleigh distribution is evaluated by the method of moments. The example of the graph of the empirical and theoretical distribution of the distances between stops and the city center in Sumy is in Figure 8.

In this case, as well as in the other nine cases, according to χ^2-test, the hypothesis about the Rayleigh distribution of distances d_j is not refuted at the significance level of 5%. The distribution parameters of d_j for all the cities are provided in Table 1. The distribution graphs are in Figure 9.

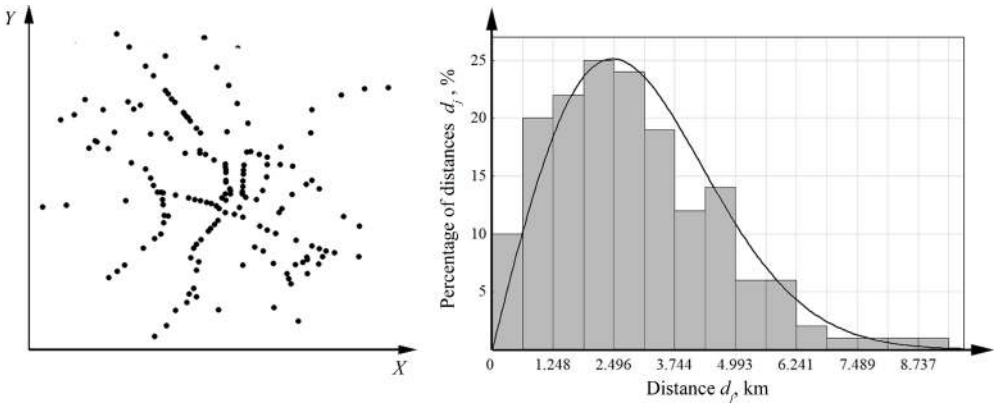


Figure 8: Stop locations and the distribution of straight-line distances between a "central" stop and the other mass transit stops in Sumy

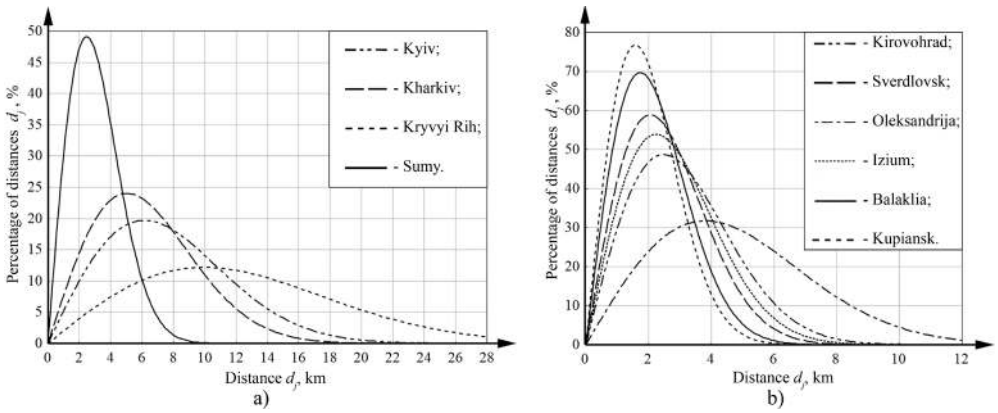


Figure 9: Distribution of straight-line distances between a "central" stop and other transit stops in Ukrainian cities: a) with population more than 250,000; b) with population under 250,000

Table 1: Parameters of the Rayleigh distribution to describe the distances between a "central" stop and the other stops

City	Index			
	Parameter σ_d	χ^2-value	Probability of χ^2-test	Expectation, km
Kyiv	6.16	33.84	.154	7.677
Kharkiv	5.05	27.59	.069	6.136
Kryvyi Rih	9.95	26.88	.081	11.717
Sumy	2.47	14.93	.312	2.907
Kirovohrad	2.49	30.13	.068	3.129
Sverdlovsk	2.06	4.88	.675	2.280
Oleksandrija	3.82	10.84	.055	4.916
Izium	2.25	25.18	.067	3.026
Balaklia	1.74	3.45	.486	2.073
Kupiansk	1.58	12.43	.087	1.879

At a first approximation, it confirms the hypothesis about the random distribution of PT stops as population and city area grow and people aspire to live and work as close as possible to the city center. The values of σ_d in Table 1 show the correlation between the transit network topology and the Rayleigh distribution parameter—the closer the city boundaries to the circumference, the lower the parameter value. In most cases, the expectation depends on the city area—the bigger the city, the greater the expectation (see Table 1).

The next step of experimental research is a validation test for the Rayleigh distribution to describe link lengths. In this case, one of the link endpoints is considered as the origin of coordinates that provides a good graphic interpretation of the relative position of adjacent stops (see Figure 10). The figure below contains a frequency histogram and the graph of the density function of the Rayleigh distribution in the city of Kharkiv.

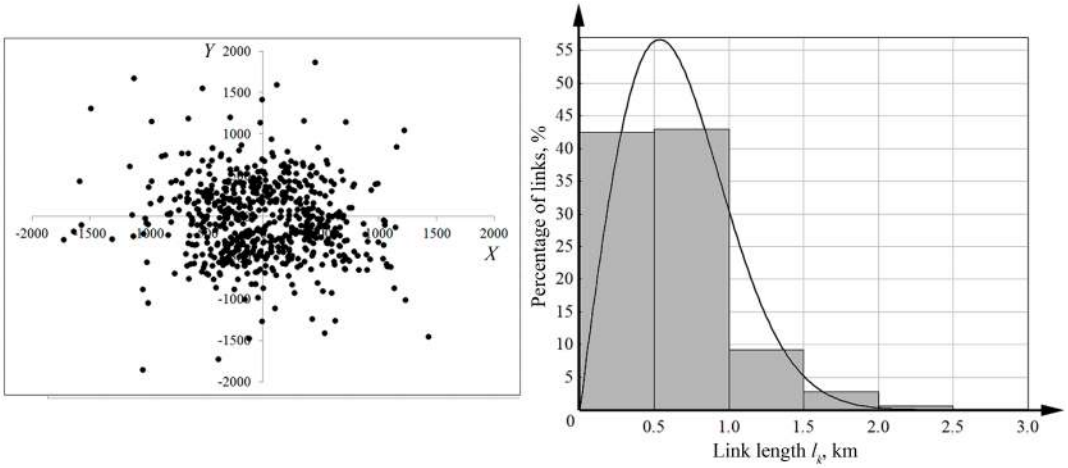


Figure 10: Relative position of link endpoints and link length distribution (Rayleigh distribution) for mass transit in the city of Kharkiv

The conformance evaluation between empirical data and theoretical curves is made using χ^2-test. The degree of conformity between empirical and theoretical distributions is higher than in the previous case, and the hypothesis is not refuted at the significance level of 5% (see Table 2).

The distribution density graphs for other cities are in Figure 11. These results confirm the hypothesis about the randomness of stop-scattering processes and indicate the suitability of using a shift parameter in the exponential distribution of link lengths.

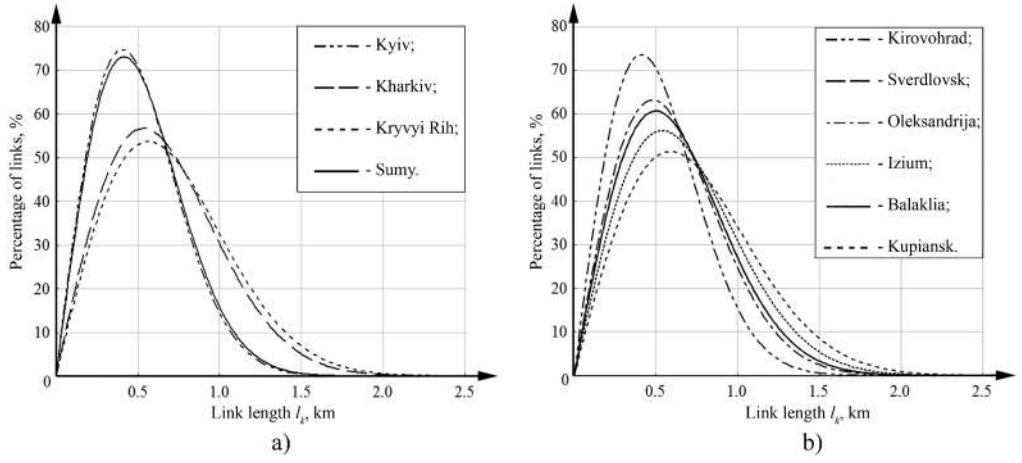


Figure 11: Graphs of the Rayleigh distribution density function for the link lengths in Ukrainian cities: a) with population over 250,000; b) with population under 250,000

Table 2: Parameters of the Rayleigh distribution to describe mass transit link lengths

City	Index			
	Parameter σ	χ^2-value	Probability of χ^2-test	Expectation, km
Kyiv	0.41	27.35	.117	0.558
Kharkiv	0.54	22.63	.179	0.686
Kryvyi Rih	0.57	24.25	.113	0.748
Sumy	0.42	16.14	.373	0.512
Kirovohrad	0.41	18.03	.115	0.494
Sverdlovsk	0.50	15.51	.973	0.638
Oleksandrija	0.48	29.22	.109	0.671
Izium	0.54	22.37	.099	0.656
Balaklia	0.50	17.53	.419	0.649
Kupiansk	0.59	20.11	.269	0.666

The data in Table 2 indicate a slight variation in σ and a low correlation between the distribution parameter and city characteristics (population, area, layout). It points to local distinctive features in transit network planning. At the same time, the expectation can be a descriptor of transit network accessibility.

The theoretical part of the paper points to the expediency of the use of the exponential distribution having shift parameter $l_{\min}$ to describe link lengths. The exponential distribution density concerning the adjusted link lengths is as follows:

$$f(l_{k\,\text{adj}}) = \lambda e^{-\lambda(l_k - l_{\min})} = \lambda e^{-\lambda \cdot l_{k\,\text{adj}}}, \quad (l_{k\,\text{adj}} = l_k - l_{\min} > 0). \tag{24}$$

An example of the Pearson test for fit of the distribution, presented by Equation 24, and the adjusted link lengths in Kyiv is in Figure 12; the distribution curves for the other cities are in Figure 13.

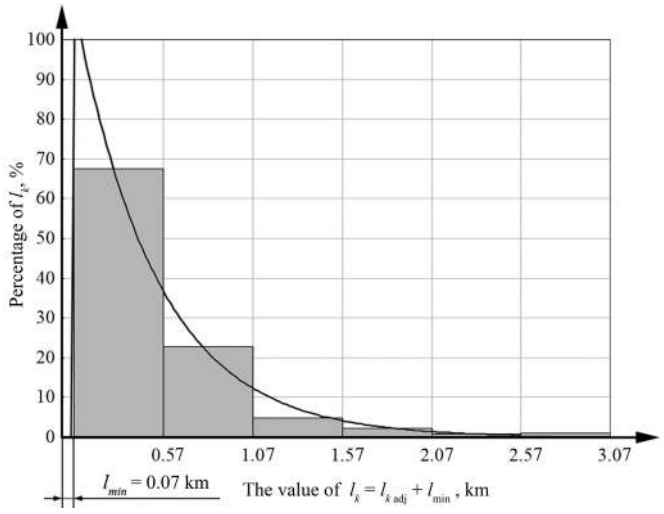


Figure 12: Exponential distribution of mass transit link lengths in Kyiv

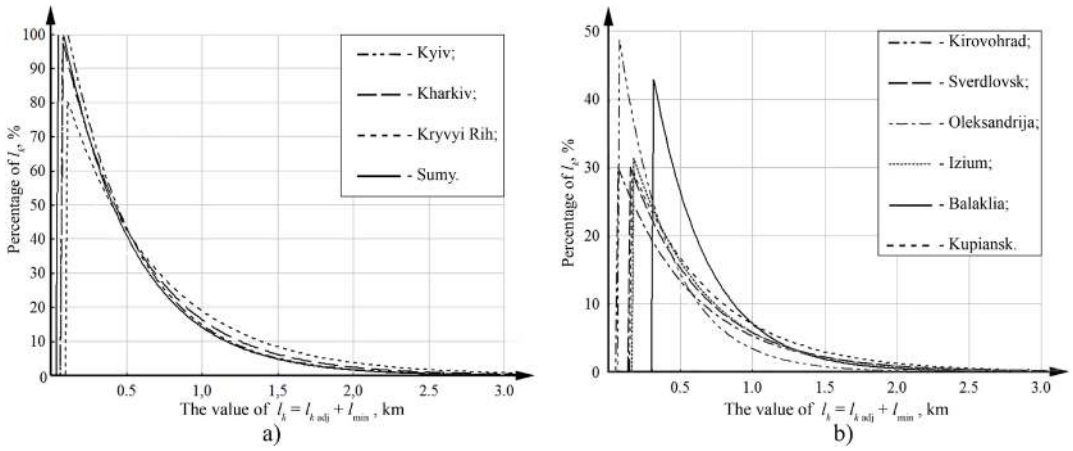


Figure 13: Exponential distribution having the shift parameter to describe mass transit link lengths in the cities: a) with population over 250,000; b) with population under 250,000

The parameter of the exponential distribution ranges from 1.63 to 2.94 (see Table 3). The degree of conformity between empirical and theoretical distributions is rather high. Only for the city of Kupiansk the probability of χ^2-test is less than 10% and equals 5.9%.

Table 3: Exponential distribution parameters to describe adjusted link lengths

City	Index			
	Minimal link length l_{min}, km	Parameter λ	χ^2-value	Probability of χ^2-test
Kyiv	0.070	2.18	2.74	.254
Kharkiv	0.074	1.96	2.74	.254
Kryvyi Rih	0.100	1.63	1.66	.436
Sumy	0.043	2.16	1.80	.773
Kirovohrad	0.077	2.94	0.60	.438
Sverdlovsk	0.147	1.95	3.71	.054
Oleksandrija	0.057	1.93	0.83	.659
Izium	0.172	2.01	6.40	.094
Balaklia	0.307	2.63	1.32	.251
Kupiansk	0.150	1.81	12.14	.059

The parameters in Table 2 and Table 3 characterize the link lengths in the cities and have no correlation with city characteristics. In this case, to estimate the transit network accessibility the parameters l_{min} and 1/λ can be used. Moreover, the use of the exponential distribution (with shift parameter l_{min}) is preferable to the use of Rayleigh distribution for link length description because the former allows us to demonstrate the special feature of a link length variable. As $l_{k\,adj} = l_k - l_{min}$, a real (full) link length is determined as $l_k = l_{k\,adj} - l_{min}$ which is purposely applied to the X-axis in Figure 12. It emphasizes that the distribution density of l_k is determined by the distribution of $l_{k\,adj}$ and it has a range starting from minimal link length l_{min} rather than 0. It completely corresponds to reality as there is no link having the zero length. In addition, when describing the link lengths, the exponential distribution has better descriptive statistics than Rayleigh one (see the probability of χ^2-test in Table 2 and Table 3).

So, we can start the evaluation of the parameters of link lengths $l_{s\,adj}$ (they constitute distances $l_{ij\,adj}$). The distribution parameters of $l_{s\,adj}$ are evaluated using a sampling method because getting the complete set of the links, which constitute the routes between all possible pairs of stops, is rather a complicated task. It is stipulated by the fact that VISUM lacks proper tools. Therefore, a random sample of 100 distances between stops was made for each city. This allowed forming the set of links that are components of all distances in the sample. The empirical frequency histograms and the graphs of theoretical distribution density functions were made (see Figure 14).

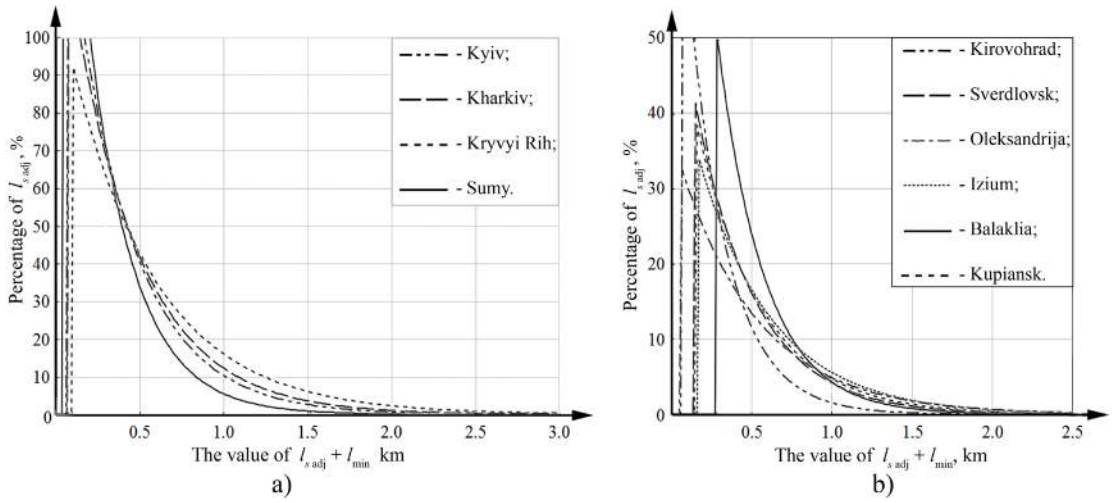


Figure 14: Exponential distribution to describe link lengths on the routes between mass transit stops in Ukrainian cities (the fact that the same links can be within various routes is taken into account): a) with population over 250,000; b) with population under 250,000

The results of the evaluation of the conformity between empirical and theoretical distributions using the χ^2-test testify that the exponential distribution with the shift parameter l_{min} is suitable to describe the adjusted link lengths $l_{s\,adj}$ in all the cities under investigation.

The parameter of the exponential distribution for $l_{s\,adj}$ is within 1.88-3.93. In all the cases, the parameter of the distribution exceeds the parameter for an initial set of link lengths $l_{k\,adj}$ (see Table 3 and Table 4). The difference in parameters ranges from 2% (the city of Izium) to 65% (the city of Sumy). So the parameters in Table 3 and Table 4 indicate that short links are more frequent than long ones along transit passenger routes and the average link length, which is traveled by passenger, is shorter than the average link length in the network.

Table 4: Parameters of the exponential distribution to describe the link length that constitute the distances between stops

City	Index		
	Parameter λ_s	χ^2-value	Probability of χ^2
Kyiv	2.66	3.65	.239
Kharkiv	2.38	2.91	.184
Kryvyi Rih	1.88	2.77	.350
Sumy	3.57	0.08	.827
Kirovohrad	3.93	7.24	.088
Sverdlovsk	2.56	2.21	.533
Oleksandrija	1.98	1.32	.569
Izium	2.05	22.11	.050
Balaklia	3.11	0.28	.627
Kupiansk	2.36	8.41	.363

It confirms the change of the exponential distribution parameter of the link lengths $l_{k\,adj}$ when they are considered to be the components ($l_{s\,adj}$) of the shortest distances between stops.

4.2 Verification of hypothesis of the gamma distribution of distances between stops

So far, the experimental verification has refuted no hypotheses. It enables us to move to the verification of the type and parameters of the distribution of the distances between stops l_{ij}. To calculate distance

matrices in VISUM, the zoning of the city areas was made and each stop was distinguished as a separate transportation zone. Having zoned the city areas, VISUM calculates cost (skim) matrices that contain the in-vehicle distances between stops in the PT system (VISUM, 2010).

To prove the hypothesis about the possibility of using a gamma distribution to describe l_{ij}, the verification test for the Poisson and triangular distribution to describe the variable n_{ij} is made. The test is performed using the χ^2 statistic. The results of the verification in the cities of Kirovohrad and Sverdlovsk are provided in Figure 15; the distribution curves for the other cities are shown in Figure 16 and Figure 17.

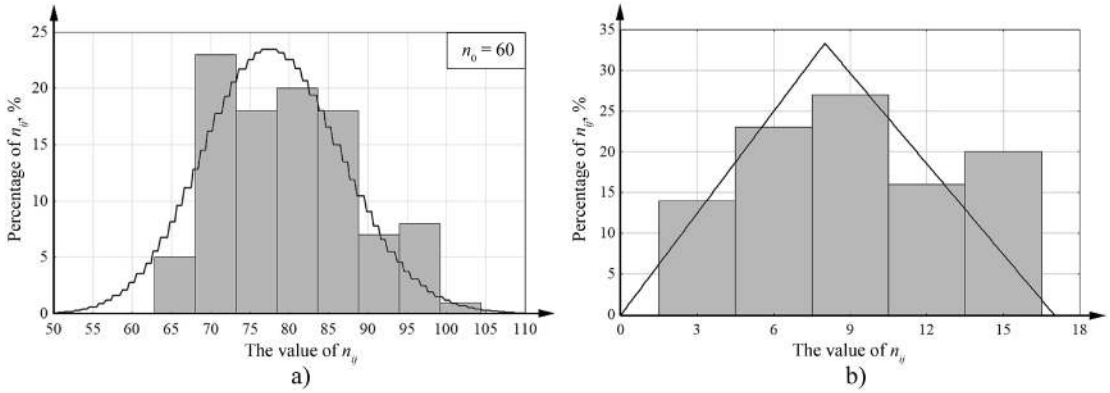


Figure 15: Distribution of the number of links between stops: a) Kirovograd, the Poisson distribution; b) Sverdlovsk, the triangular distribution

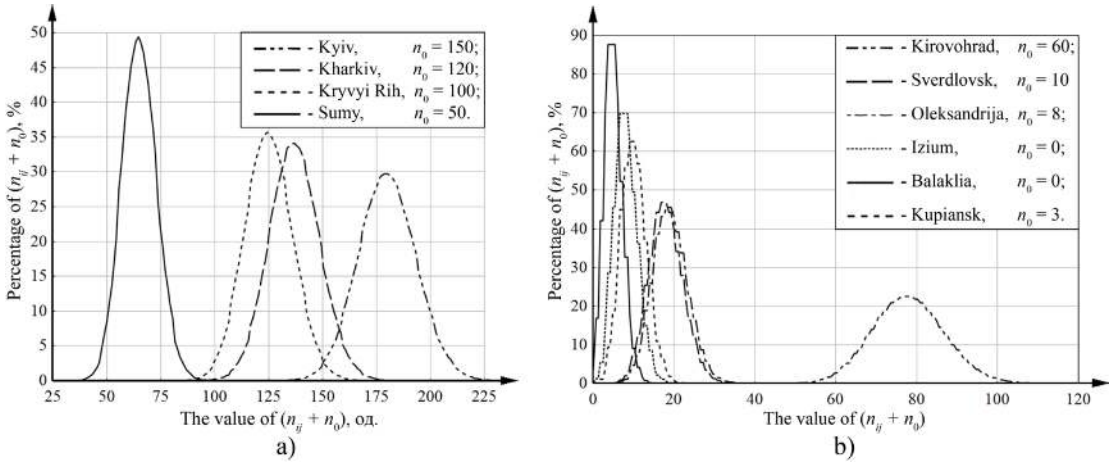


Figure 16: The Poisson distribution of the number of links between stops in the Ukrainian cities: a) with population over 250,000; b) with population under 250,000

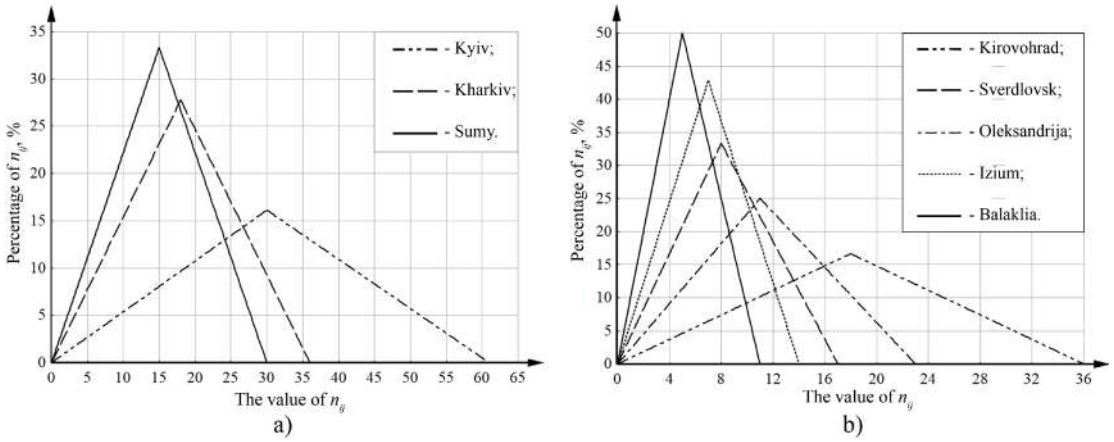


Figure 17: The triangular distribution of the number of links between stops in Ukrainian cities: a) with population over 250,000; b) with population under 250,000

The best distribution to describe the n_{ij} is taken into account for two types of the triangular distribution.

Of 10 cases, the Poisson distribution is unsuitable to describe the number of stops in the city of Kryvyi Rih only. The triangular distribution is unsuitable in two cities—Kryvyi Rih and Kupiansk (see Table 5 and Table 6). It can be regarded as a consequence of poor theoretical validation of the n_{ij} distribution.

Table 5: Parameters of the Poisson distribution of the number of links between stops

City	Index			
	Distribution parameter	Shift parameter $n_0 = const$	χ^2-value	Probability of χ^2
Kyiv	180	150	2.96	.889
Kharkiv	137	120	14.39	.072
Sumy	65	50	6.10	.192
Kirovohrad	78	60	6.52	.259
Sverdlovsk	18	10	5.68	.460
Oleksandrija	19	8	12.43	.053
Izium	8	0	8.35	.214
Balaklia	5	0	6.32	.097
Kupiansk	10	3	3.00	.392

Table 6: Parameters of the triangular distribution of the number of links between stops

City	Index			
	Distribution parameter	Even or odd maximal value	χ^2-value	Probability of χ^2
Kyiv	30	Even	8.29	.600
Kharkiv	17	Odd	23.85	.068
Sumy	14	Odd	17.29	.068
Kirovohrad	17	Odd	18.09	.054
Sverdlovsk	8	Even	8.53	.074
Oleksandrija	11	Even	16.50	.057
Izium	6	Odd	17.22	.102
Balaklia	5	Even	14.35	.073

The parameters of the Poisson and triangular distribution presented in Table 5 and Table 6 highlight a general tendency of the decreasing of the parameter while the area decreases.

In general, the correspondence between the empirical distribution of n_{ij} and theoretical one can be

considered sufficient to move on to the verification of conformity between the distribution of distances between stops and the gamma distribution. The example of the verification results in the city of Kryvyi Rih is shown in Figure 18.

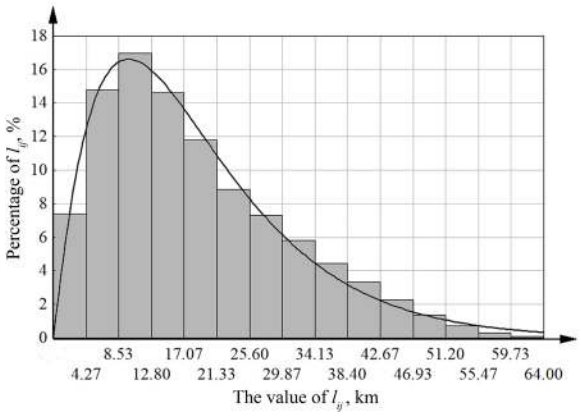


Figure 18: The gamma distribution of distances between stops in Kryvyi Rih

The results of the verification are analogous in the other cities; the distribution curves are shown in Figure 19. The distribution parameters are in Table 7. The probability of the χ^2-test shows that the gamma distribution can be used to describe distances between stops.

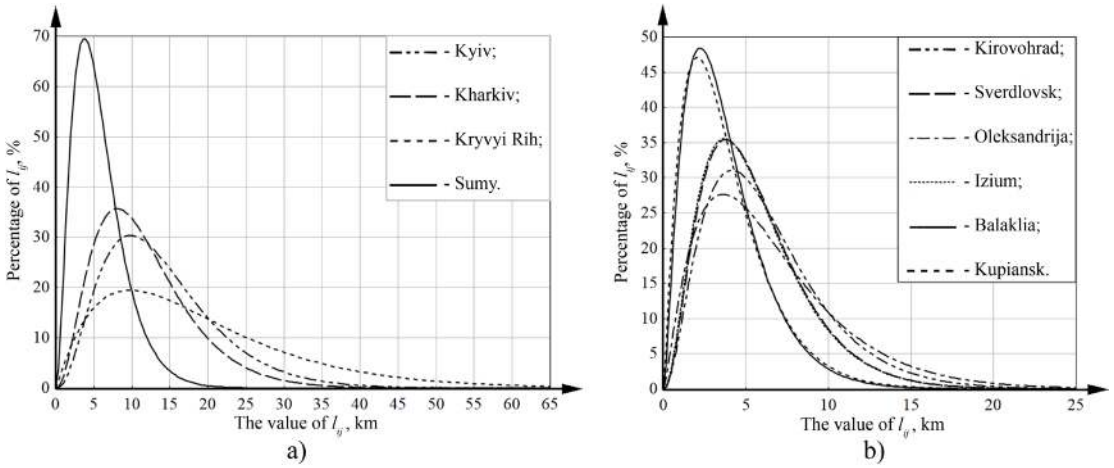


Figure 19: The gamma distribution of distances between stops in the cities: a) with population over 250,000; b) with population under 250,000

Table 7: Parameters of the gamma distribution of distances between transit stops

City	Index			
	Scale parameter	Shape parameter	χ^2-value	Probability of χ^2
Kyiv	4.10	3.40	21.14	.145
Kharkiv	3.59	3.25	16.45	.280
Kryvyi Rih	9.18	2.07	8.04	.582
Sumy	2.00	2.88	7.16	.343
Kirovohrad	2.11	2.96	8.96	.439
Sverdlovsk	1.79	3.11	26.25	.050
Oleksandrija	2.91	2.24	15.98	.091
Izium	1.81	3.04	28.11	.051
Balaklia	1.57	2.41	20.09	.063
Kupiansk	1.76	2.16	6.61	.671

The data in Table 7 show a poor correlation between the distribution parameters of the distances between transit stops and the city area, population or layout. It emphasizes that the fundamentals developed can be put into practice ignoring local PT features.

So, the theoretical distribution that describes distances between stops with sufficient accuracy is the gamma distribution having the shape parameter greater than 2. On the basis of that, it can be concluded that the obtained gamma distribution is the result of the spatial location of trip attractors (PT stops). This fact indicates that the TLD is stipulated by the stops location in the city. It gives every reason to believe that the gamma distribution of distances between stops is the basis of the TLD function for most cities.

4.3 Experimental research of transformation from the distribution of distances between stops to the TLD

In order to estimate the influence of the O-D matrix on the transformation from the distribution of distances between stops to the TLD, it is necessary to perform the transformation using actual O-D matrices for the cities selected for experimental research. The lack of actual O-D matrices for any Ukrainian city makes the obtaining of an actual TLD function impossible. Therefore, it is expedient to research the trip length distribution in general, using theoretically possible O-D matrices. To calculate the matrices, the next trip distribution models are selected:

1. The model of trip distribution that minimizes the total distance traveled when making HBW trips in the transit system. This model is called "General Minimization." The O-D matrix is a solution for the Hitchcock-Koopmans transportation problem (Rao, 2009) having given total trip ends. It has to characterize an "extreme" state of transport demand in a theoretically possible case of "ideal" settlement when passenger transportation cost and transit operator cost are minimal.
2. A gravity model having the deterrence function $c(l_{ij})=1/l_{ij}^2$, which is called the "Classic Gravity Model."
3. A gravity model having the deterrence function $c(l_{ij})=1/l_{ij}$, which is called the "Reduced Gravity Model."
4. The model of the random filling of the O-D matrix, which is called the "Random Matrix." To apply the model, a computer program has been designed. Its operation is based on the use of random numbers to select cells and to assign the number of trips to those cells.
5. The model of trip distribution that maximizes the total distance traveled when making HBW trips in the transit system. This model is called the "General Maximization." This strategy is introduced in contrast to the "General Minimization" and it is purely theoretical. It has to demonstrate the maximum influence of the O-D matrix on the TLD function curve under given total trip ends.

Initial information to calculate the O-D matrix is the total trip ends and distance matrices in the cities under investigation. On this basis, the O-D matrices are obtained for each city in accordance with the above-mentioned models. As a result of the transformation of the distance matrices using the estimated O-D matrices, the appropriate theoretically possible sets of trip lengths are received. Evaluation of conformity between the TLD and the theoretical gamma distribution is performed using the χ^2-test. The example of the evaluation for the "Reduced Gravity Model" applied in Kirovohrad is shown in Figure 20; the other curves for Kirovohrad are in Figure 21. The results of transformations for the other cities are analogous to those for Kirovohrad (see Figure 22). This enables us to draw some conclusions.

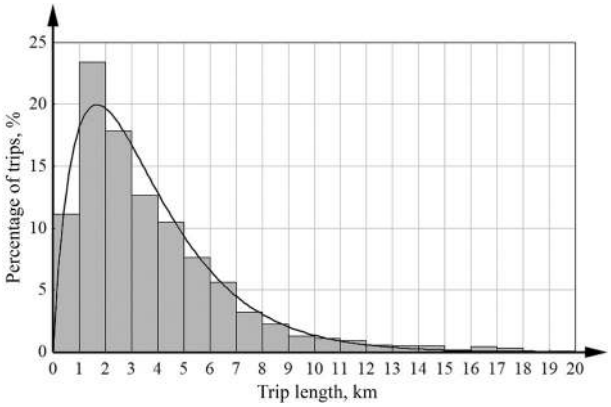


Figure 20: Evaluation of conformity between TLD and gamma distribution function when using the "Reduced Gravity Model" to calculate the O-D matrix for Kirovograd

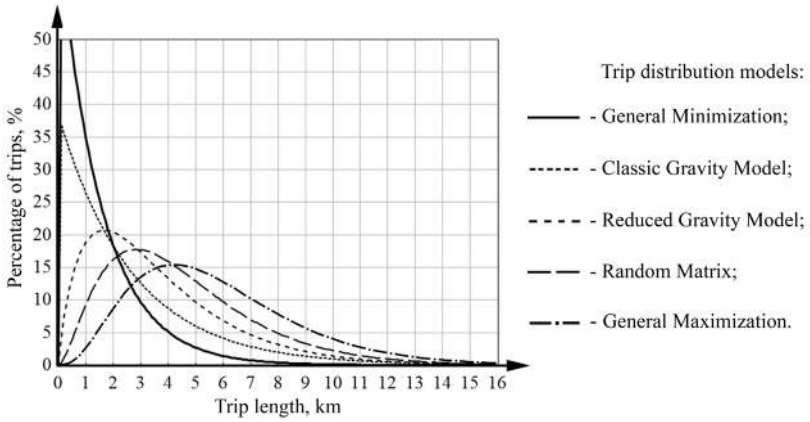


Figure 21: Trip length distributions for Kirovograd resulting from different trip distribution models

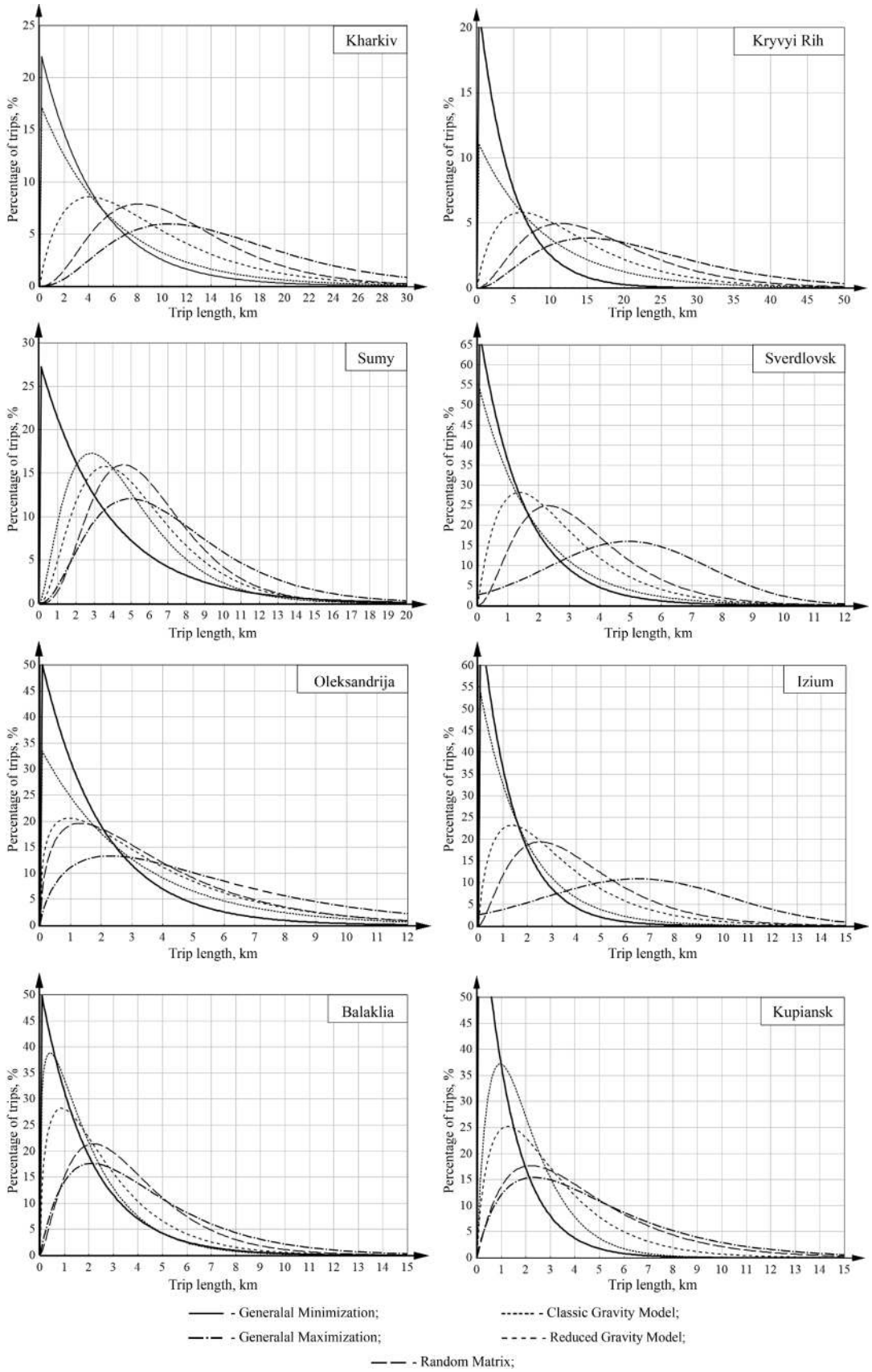


Figure 22: The TLD curves for Ukrainian cities resulting from different trip distribution models

When fitting theoretical distribution to TLDs derived from the "extreme" trip distribution models ("General Minimization" and "General Maximization") the low probabilities of the χ^2-test (less than 5%) are obtained. At the same time, all the "middle" models provide a comparatively high probability of the χ^2-test (see Table 8).

Table 8: The parameters of distributions that describe theoretically possible TLD curves in the Ukrainian cities

Trip distribution model	Distribution type	Distribution parameters		χ^2-value	Probability of χ^2
		Scale parameter	Shape parameter		
Kharkiv					
General Minimization	Exponential	0.22		25.68	.050
Classic Gravity Model	Exponential	0.17		14.67	.224
Reduced Gravity Model	Gamma	4.72	1.87	14.44	.145
Random Matrix	Gamma	3.19	3.53	14.58	.284
General Maximization	Gamma	4.24	3.50	18.69	.053
Kryvyi Rih					
General Minimization	Exponential	0.22		4.39	.280
Classic Gravity Model	Exponential	0.11		9.72	.286
Reduced Gravity Model	Gamma	6.92	1.88	4.69	.624
Random Matrix	Gamma	5.61	3.03	17.29	.066
General Maximization	Gamma	7.12	3.11	35.52	.051
Sumy					
General Minimization	Exponential	0.27		23.87	.051
Classic Gravity Model	Gamma	1.81	2.57	5.25	.425
Reduced Gravity Model	Gamma	1.75	3.07	8.38	.289
Random Matrix	Gamma	1.39	4.30	8.55	.186
General Maximization	Gamma	2.20	3.26	19.78	.052
Kirovohrad					
General Minimization	Exponential	0.64		18.08	.054
Classic Gravity Model	Exponential	0.37		7.60	.389
Reduced Gravity Model	Gamma	1.97	1.85	7.44	.397
Random Matrix	Gamma	1.73	2.64	4.97	.378
General Maximization	Gamma	1.61	3.61	8.53	.115
Sverdlovsk					
General Minimization	Exponential	0.69		5.68	.280
Classic Gravity Model	Exponential	0.54		5.02	.522
Reduced Gravity Model	Gamma	1.34	2.02	11.16	.101
Random Matrix	Gamma	1.10	3.11	2.54	.400
General Maximization	Normal	6.68	4.93	5.19	.226
Oleksandrija					
General Minimization	Exponential	0.50		3.18	.116
Classic Gravity Model	Exponential	0.33		14.93	.056
Reduced Gravity Model	Gamma	2.78	1.34	26.81	.051
Random Matrix	Gamma	2.52	1.52	8.99	.098
General Maximization	Gamma	3.32	1.69	37.85	.050
Izium					
General Minimization	Exponential	0.72		5.66	.061
Classic Gravity Model	Exponential	0.54		6.80	.188
Reduced Gravity Model	Gamma	1.87	1.73	23.67	.054
Random Matrix	Gamma	1.62	2.56	25.17	.052
General Maximization	Normal	6.49	3.80	1.47	.454
Balaklia					
General Minimization	Exponential	0.50		41.84	.050
Classic Gravity Model	Gamma	1.62	1.25	21.24	.054
Reduced Gravity Model	Gamma	1.79	1.49	11.68	.331
Random Matrix	Gamma	1.52	2.44	38.99	.053
General Maximization	Gamma	2.19	1.97	2.30	.051
Kupiansk					
General Minimization	Exponential	0.76		7.76	.479
Classic Gravity Model	Gamma	1.08	1.88	6.16	.068
Reduced Gravity Model	Gamma	1.71	1.74	11.83	.054
Random Matrix	Gamma	2.12	2.04	13.12	.098
General Maximization	Gamma	2.58	1.90	85.00	.050

The obtained plots (see Figure 20, Figure 21, and Figure 22) make it clear that any trip distribution model causes changes in the distribution of distances between stops that preserve the similarity of curves with the density function graphs of the gamma distribution family. The models "General Minimization" and "Classic Gravity Model" result in the TLD curves that are close to exponential. The "General Maximization" model results in the TLDs that can be described by the gamma distribution that is close to normal. It explains variations in the shape parameter of the gamma distribution for various cities as well as specifies a wider range for this parameter. However, the strategies of "General Minimization" and "General Maximization" can be regarded as generating low-probability variants of the O-D matrix, i.e., "extreme" transport demand states.

More realistic trip length distributions are obtained when using the "Reduced Gravity Model" and "Random Matrix" model. The result of the "Random Matrix" model is especially important as it refutes the determinism of the trip distribution process.

On the basis of the research, we can state that the gamma distribution, which describes the TLD curve, is determined by the distribution of distances between trip attractors in the city. This means that the TLD regularities are the result of the distribution of trip attractors in the direction from the city center towards its suburbs as the city expands. Besides that, it should be noted that the gamma distribution is not transformed into any other distribution irrespective of a way to distribute trips in the O-D matrix. It can be explained by constraints on total trip ends for transportation zones. The verification of this statement is the subject of separate research.

The research results call for the application of actual TLD functions to estimate the transport demand of PT users. It will allow us to validate the distribution of trips among "origin-destination" pairs in the most effective manner.

6 Conclusions

This paper states that trip length distribution is stipulated by the O-D matrix and the matrix of distances between transport attractors, which are PT stops. Theoretical background and our experiments clearly demonstrate that a typical TLD plot is conditioned by the distribution of distances between stops. The distribution of the distances is a consequence of the random location of transport attractors within the city when the city population and area increase. Herewith it is justified that the gamma distribution is suitable to describe both the regularities in distances between stops and in-vehicle trip lengths.

During the research, it was determined that the O-D matrix is a kind of the multiplicator of distances between stops to transform them into trip lengths. Though such an impact of the O-D matrix causes the change of distribution parameters only and does not change the type of the distribution of distances between stops. All the trip distribution models, including those which generate "extreme" transport demand states, cause the TLD that can be described by density functions originating from the family of the gamma distribution. This fact is a considerable contribution to the development of the theory of city population settlement patterns as it clarifies the reasons of similar TLD plots in different cities. Despite the influence of various factors, the TLD is determined by the location of trip attractors in the city that are the stops in terms of mass transit.

Also we should point that all the way to get the TLD on the base of stop coordinates and link lengths has revealed two new processes of the random variable forming. They mean a multiple repetition of each value of an initial random variable according to certain rules, which can be defined as multiplicators. The first multiplicator is represented by the O-D matrix, the second one—by the sampling of the links when forming the shortest routes between pairs of PT stops in the city. Both multiplicators are unique and do not change the type of the distribution of an initial random variable. In general, the

multiplicator can have any form including an analytical equation and it can influence on the type of initial distribution in different ways.

One more issue that deserves attention is that proven expediency of using the gamma distribution to describe the TLD function facilitates getting factual settlement regularities. Having determined the type of distribution, it is sufficient to conduct a selective survey to estimate distribution moments only. In terms of practical application, it considerably decreases the cost of getting the TLD function.

The results of this research make a contribution to the methodology of trip distribution. A typical TLD plot, which is produced by a randomly filled O-D matrix, points to the fact that the use of the standard gravity and entropy models for transportation planning is not compulsory. The classical models can be substituted with more prospective models that are constrained by both the total trip ends and the TLD function. The latter models can be a subject for further research as they allow producing O-D matrices that completely correspond to the TLD function without any additional calibration. Another problem to be investigated is to apply the theoretical fundamentals of the TLD of HBW trips for other transportation modes and trip types.

References

Acheampong, R. A., & Silva, E. A. (2015). Land use-transport interaction modeling: A review of the literature and future research directions. *Journal of Transport and Land Use, 8*(3), 11–38.

Ahern, A., Weyman G., Redelbach M., Schulz A., Akkermans L., Vannacci L., Anoyrkati, E., & van Grinsven, A. (2013). *Analysis of national travel statistics in Europe OPTIMISM WP2: Harmonization of national travel statistics in Europe* (Report EUR 26054 EN). Retrieved from http://publications.jrc.ec.europa.eu/repository/bitstream/JRC83304/tch-d2.1_final.pdf

Aultman-Hall, L., Sears J., Dowds J., & Hines P. (2012). *Spatial analysis of travel demand and accessibility in Vermont: Where will EVs work?* (TRC Report 12-007) Retrieved from http://www.uvm.edu/~transctr/research/trc_reports/UVM-TRC-12-007.pdf

Benson, J. D., Teniente M. F., Stover, V. G., & Cunagin, W. D. (1979). *An improved model for the estimation of trip length frequency distribution* (Report No. 0194-5). Retrieved from https://library.ctr.utexas.edu/digitized/texasarchive/80-2-7.pdf

Bovy, P. H. L., Bliemer, M. C. J., & van Nes, R. (2006). *Transportation modeling.* Delft, Netherlands: Delft University of Technology.

Chow, Y. S., & Teiher, H. (1978). *Probability theory.* New York, NY: Springer-Verlag.

Cibulka, J. (1987). Качество пассажирских перевозок в городах [*Quality of city passenger transportation*]. Moscow: Transport.

Cox, D. R., & Smith, W. L. (1961). *Queues*. Hoboken, NJ: Wiley.

Daniels, R., & Mulley, C. (2013). Explaining walking distance to public transport: The dominance of public transport supply. *Journal of Transport and Land Use, 6*(2), 5–20.

Efremov, I. S., & Golc G. A. (1988). Городской пассажирский транспорт и АСУ транспорта [*City public transport and automatic control systems in transportation*], in V. M. Kobozev (Ed.). Moscow: Nauka.

El-Geneidy, A., Grimsrud, M., Wasfi, R., Tétreault, P., & Surprenant-Legault, J. (2014). New evidence on walking distances to transit stops: Identifying redundancies and gaps using variable service areas. *Transportation, 41*(1), 193–210.

Englund, D., Eash, R., & Lupa, M. (2010, June). *Matching workers and employment opportunities: Linking worker incomes and wages in regional travel models.* Paper presented at the Transport Chicago conference, Chicago, IL. Retrieved from http://www.transportchicago.org/uploads/5/7/2/0/5720074/matching_workers_and_ employment_opportunities.pdf

Feller, W. (1966). *An introduction to probability theory and its applications* (Vol. II). Hoboken, NJ: Wiley.

Forbes, C., Evans, M., Hastings, N., & Peacock, B. (2011). *Statistical distributions* (4th ed.). Hoboken, NJ: Wiley.

Fricker, J. D., & Jin, L. (2008). *Development of an integrated land-use transportation model for Indiana* (Final Report FHWA/IN/JTRP-2008/15). West Lafayette, IN: Purdue University, School of Civil Engineering.

Gehrke, S. R., & Clifton, K. J. (2016). Toward a spatial-temporal measure of land-use mix. *Journal of Transport and Land Use, 9*(1), 171–186.

Harznagy, A., Fi, I., London, A., & Neŧmeth, T. (2015, June). *Complex network analysis of public transportation networks: A comprehensive study.* Paper presented at the 2015 Models and Technologies for Intelligent Transportation Systems (MT-ITS) conference, Budapest, Hungary. Abstract retrieved from http://ieeexplore.ieee.org/document/7223282/.

Horner, M. W., & Downs, J. A. (2014). Integrating people and place: A density-based measure for assessing accessibility to opportunities. *Journal of Transport and Land Use, 5*(2), 23–40.

Huntsinger, L. F., & Donnelly, R. (2014, January). *Reconciliation of regional travel model and passive device tracking data.* Paper presented at the Transportation Research Board 93rd Annual Meeting, Washington, DC.

Junge, J. R., & Levinson, D. (2012). Prospects for transportation utility fees. *Journal of Transport and Land Use, 5*(1), 33–47.

Katsis, P., Papageorgiou, T., & Ntziachristos, L. (2014). Modelling the trip length distribution impact on the CO2 emissions of electrified vehicles. *Energy and Power, 4*(1A), 57–64.

Kim, J., & Lee, S. (2001). A gradient method for the estimation of travel demand using traffic counts on the large scale network. *Journal of the Eastern Asia Society for Transportation Studies, 4*(2), 351–357.

L & A Transportation (2006). *Lincoln MPO travel demand model* (Model documentation). Princeton, NJ: Lima & Associates Transportation—G.I.S.

Milakis, D., Cervero, R., & Wee, L. B. (2015). Stay local or go regional? Urban form effects on vehicle use at different spatial scales: A theoretical concept and its application to the San Francisco Bay Area. *Journal of Transport and Land Use, 8*(2), 59–86.

Moeckel, R. (2017). Constraints in household relocation: Modeling land-use/transport interactions that respect time and monetary budgets. *Journal of Transport and Land Use, 10*(1), 211–228.

Mounir, M. M. A.-A. (2014). Calibrating a trip distribution gravity model stratified by the trip purposes for the city of Alexandria. *Alexandria Engineering Journal, 53*, 677–689.

Ortuzar, J. D., & Willumsen, L. G. (2011). *Modelling transport* (4th ed.). Chichester, England: John Wiley & Sons.

O'Sullivan, S., & Morrall, J. (1996). Walking distances to and from light-rail transit stations. *Transportation Research Record, 1538*, 19–26.

Porter, C. D., Brown, A., Dunphy, R. T., & Vimmerstedt, L. (2013). *Effects of the built environment on transportation: Energy use, greenhouse gas emissions, and other factors* (Research Report). Retrieved from https://www.nrel.gov/docs/fy13osti/55634.pdf

Rao, S. S. (2009). *Engineering optimization: Theory and practice* (4th ed.). Hoboken, NJ: Wiley.

Riordan, J. (1962). *Stochastic service systems.* Hoboken, NJ: Wiley.

Saaty, T. L. (1961). *Elements of queueing theory.* New York: McGraw-Hill.

Saff, E. B., & Snider, A. D. (1976). *Fundamentals of complex analysis for mathematics, science and engineering.* Englewood Cliffs, NJ: Prentice Hall.

Shelejhovskij, G. V. (1946). Композиция городского плана как проблема транспорта [*City layout as a problem of transport*]. Moscow: GIPROGOR.

Srinivasan, S., Provost, R., & Steiner, R. (2013). Modeling the land-use correlates of vehicle-trip lengths for assessing the transportation impacts of land developments. *Journal of Transport and Land Use, 6*(2), 59–75.

Stead, D., & Marshall, S. (2001). The relationships between urban form and travel patterns. An international review and evaluation. *European Journal of Transport and Infrastructure Research, 1*(2), 113–141.

Transportation Research Board. (2010). *NCHRP synthesis 406: Advanced practices in travel forecasting. A synthesis of highway practice* (Report). Retrieved from http://onlinepubs.trb.org/onlinepubs/nchrp/nchrp_syn_406.pdf

Transportation Research Board. (2013). *Transit capacity and quality of service manual* (TCRP report 165). Retrieved from http://www.trb.org/Main/Blurbs/169437.aspx

Veenstra, S. A., Thomas, T., & Tutert, S. I. A. (2010). Trip distribution for limited destinations: A case study for grocery shopping trips in the Netherlands. *Transportation, 37*, 663–676.

VISUM (Version 10.0). (2010). Karlsruhe, Germany: PTV Planung Transport Verkehr AG.

Wegener, M., & Fuerst, F. (2004). Land-use transport interaction: State of the art. *SSRN Electronic Journal*. doi:10.2139/ssrn.1434678

Xie, F., & Levinson, D. (2011). *Evolving transportation networks.* New York: Springer.

Yang, F., Jin, P. J., Wan, X., Li, R., & Ran, B. (2013, January). *Dynamic origin-destination travel demand estimation using location-based social networking data.* Paper presented at the Transportation Research Board 92nd Annual Meeting, Washington, DC.

Yigitcanlar, T., Dodson, J., Gleeson, B., & Sipe, N. (2005). *Sustainable Australia: Containing travel in master planned estates.* Brisbane, Australia: Griffith University, URP.

Zhao, F., Chow, L.-F., Li, M.-T., & Gan, A. (2014). *Refinement of FSUTMS trip distribution methodology* (Final Report for BB942). Retrieved from http://lctr.eng.fiu.edu/re-project-link/bb942.pdf

Appendix A: Clarification of relationship between Rayleigh distribution and exponential distribution having the shift parameter when describing link lengths

The conclusion about the Rayleigh distribution of the link lengths seems to contradict the exponential distribution of $l_{k\min}$. However, the contradiction can be resolved by the property of the Rayleigh distribution which reads that it corresponds to the squared exponential variable. Therefore, we should replace a squared link length with a linear one. To do it we can apply the linearization using expansion in a Taylor series (Saff & Snider, 1976):

$$l_k^2 = v(l_k) = v(t) + v'(t) \cdot (l_k - t), \tag{A1}$$

where:

t Point at which expansion is made

$v'(t)$ First derivative of the function under investigation at a point t; in this case, the function under investigation is l^2_k and its derivative is $v'(l_k)=(l^2_k)'=2l_k$

For Taylor expansion it is expedient to choose the point 0.5 km at which the constant $v'(l_k)=1$. However, there will be an actual link length shift Δl =0.25 km:

$$l_k^2 = l_k - \Delta l = l_{k\,\text{adj}} \tag{A2}$$

where:

$l_{k\,\text{adj}}$ Random component of the k-th link, km

Then under the condition $\lambda = 1/(2\sigma^2)$, the Rayleigh distribution becomes an exponential one:

$$F(l_{k\,\text{adj}}) = 1 - e^{-\lambda(l_k - \Delta l)} = 1 - e^{-\lambda l_{k\,\text{adj}}}, \tag{A3}$$

where:

$$l_{k\,\text{adj}} = (l_k - \Delta l) \geq 0. \tag{A4}$$

It is essential that the linearization of a squared link length leads to the appearance of a constant component Δl in Equation A24 which is the shift parameter of link length distribution. So, after link length shortening by constant Δl, the random components of the link lengths $l_{k\,\text{adj}}$ should correspond to exponential distribution. The result is analogous to Equation 6 where $l_{k\min} \leq l_k$ corresponds to the exponential distribution.

Therefore, the next step to experimentally verify the results is to verify the exponential distribution of the random components of link length $l_{k\,\text{adj}}$. To do that, it is necessary to determine the corresponding shift parameter Δl. Taking into account the properties of the exponential distribution, it is reasonable to use the minimal link length $l_{\min}$ as a shift parameter in each city:

$$\Delta l = \min(\forall l_k, k \in \{1, K\}) = l_{\min}, \tag{A5}$$

where:

K Total number of links in the city

Appendix B: Justification of the exponential distribution of the link lengths

To determine the theoretical presupposition of the distribution law of the variable $l_{k\,\mathrm{adj}}$ it is expedient to consider other options for the formation of link length that differ from those mentioned in the theoretical part of the paper. The much more realistic scenario is the setting up of several intermediate stops between the existing final stop in the PT system and a new final stop in the expanded city area. From a mathematical point of view, the scenario can be represented as the random location of stops between those two last stops within a given segment (see Figure B1).

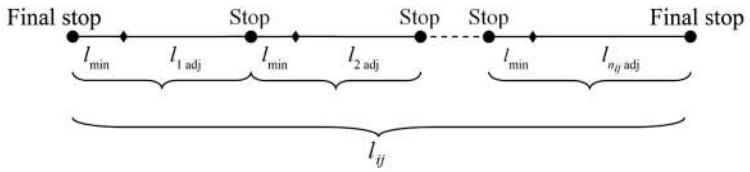


Figure B1: Graphical interpretation of distance l_{ij} between stops *i* and *j*

If the sum $\sum_{k=1}^{n_{ij}} l_{k\,\mathrm{adj}}$ in Equation 11 is designated as $l_{ij\,\mathrm{adj}}$, then it can be written as follows:

$$l_{ij\,\mathrm{adj}} = \sum_{k=1}^{n_{ij}} l_{k\,\mathrm{adj}} = l_{ij} - n_{ij} \cdot l_{\min} = \sum_{k=1}^{n_{ij}} (l_k - l_{\min}). \tag{B1}$$

So $l_{il\,\mathrm{adj}}$ will be the sum of the random components of the link lengths which complete the $l_{\min}$ to make up l_k. Hence $l_{k\,\mathrm{adj}}$ can be the result of a random division of $l_{il\,\mathrm{adj}}$ by n_{ij} random components (see Figure B2).

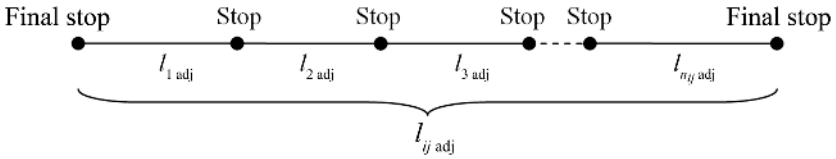


Figure B2: Division of the $l_{ij\,\mathrm{adj}}$, which is the sum of the adjusted link lengths $l_{k\,\mathrm{adj}}$, into components

The difference between Figure B1 and Figure B2 is in the superposition of the endpoints of the segments $l_{\min}$ that is identical to their removal. If we consider addends $l_{1\,\mathrm{adj}}$, $l_{1\,\mathrm{adj}} + l_{2\,\mathrm{adj}}, \ldots, l_{1\,\mathrm{adj}} + l_{2\,\mathrm{adj}} + \ldots + l_{(n_{ij}-1)\,\mathrm{adj}}$ referring to Figure 24, they are order statistics $l_{1\,\mathrm{adj}} = x_{(1)}$, $l_{1\,\mathrm{adj}} + l_{2\,\mathrm{adj}} = x_{(2)}, \ldots, l_{1\,\mathrm{adj}} + l_{2\,\mathrm{adj}} + \ldots + l_{(n_{ij}-1)\,\mathrm{adj}} = x_{(n_{ij}-1)}$ of the random variable *X* that are uniformly distributed on the segment $l_{ij\,\mathrm{adj}}$ and observed in $(n_{ij}-1)$ events—the experiments which result in the appearance of stops on the route from *i* to *j*. The number of experiments (the events of the appearance of stops on the route from *i* to *j*) is $n = (n_{ij} - 1)$ and it is Poisson distributed:

$$P_n(l_{ij\,\mathrm{adj}}) = \frac{(\mu \cdot l_{ij\,\mathrm{adj}})^n}{n!} e^{-\mu \cdot l_{ij\,\mathrm{adj}}}, \quad \mu = \frac{1}{\bar{n}}, \quad n = n_{ij} - 1, \tag{B2}$$

where:

μ Poisson distribution parameter

$\bar{n}$ Mean of the number of stops on the route between *i* and *j*

The $l_{ij\,\mathrm{adj}}$ can be investigated in two processes: 1) the process of the appearance of *n* stops that are

uniformly distributed within the segment (0, $l_{ij\,\mathrm{adj}}$); 2) the Poisson process where the distribution presented by Equation B2 is well known.

It is expedient to introduce random event A_n , which means that within intervals $\{\beta_i\}$ there will be one stop only (the point of the Poisson process), and within intervals $\{\alpha_i\}$ there will be no stop (see Figure B3).

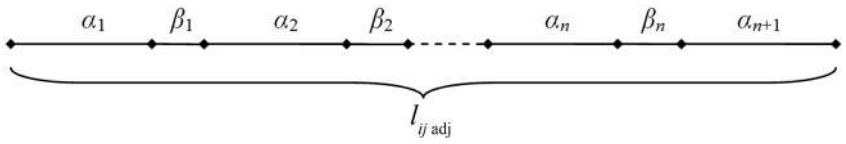


Figure B3: Graphic representation of event A_n that implies the location of one stop within intervals $\{\beta_i\}$ and no stop within intervals $\{\alpha_i\}$

For the appearance of *n* uniformly distributed stops within the segment (0, $l_{ij\,\mathrm{adj}}$) the probability A_n is determined as follows:

$$P\{A_n|B\}=\left(\frac{\beta_1}{l_{ij\,\mathrm{adj}}}\right)\cdot\left(\frac{\beta_2}{l_{ij\,\mathrm{adj}}}\right)\cdot\ldots\cdot\left(\frac{\beta_n}{l_{ij\,\mathrm{adj}}}\right)\cdot n!=\frac{\beta_1\beta_{2\ldots}\beta_n}{(l_{ij\,\mathrm{adj}})^n}\cdot n!, \tag{B3}$$

where:

B Event that there will appear *n* stops within the segment (0, $l_{ij\,\mathrm{adj}}$)

Multiplier *n!* is explained by the fact that there is no difference when *n* points are repositioned on *n* intervals (Chow & Teiher, 1978).

For the Poisson process (Feller, 1966)

$$P\{A_n|B\}=\frac{P\{A_n\cdot B\}}{P\{B\}}. \tag{B4}$$

As for the Poisson flow, the events on non-overlapping intervals are independent (Riordan, 1962), the probability of a joint event $\{A_n \cdot B\}$ can be calculated as the product of the probabilities of separate events

$$P(C)=\mu\cdot\beta_i\cdot e^{-\mu\beta_i},\quad P(D)=e^{-\mu\alpha_i}, \tag{B5}$$

where:

C Event when one stop appears within an interval β_i

D Event when no stop appears within an interval α_i

The unconditional probability of event *B* is determined by Equation B2. It follows that

$$P\{A_n|B\}=\left(\mu\beta_1\cdot\mu\beta_2\ldots\mu\beta_n\cdot e^{-\mu\beta_1}\cdot e^{-\mu\beta_2}\ldots e^{-\mu\beta_n}\cdot e^{-\mu\alpha_1}\cdot e^{-\mu\alpha_2}\ldots e^{-\mu\alpha_{n+1}}\right)\Big/\frac{(\mu\cdot l_{ij\,\mathrm{adj}})^n}{n!}e^{-\mu\cdot l_{ij\,\mathrm{adj}}}=\\ =\frac{\beta_1\beta_{2\ldots}\beta_n}{(l_{ij\,\mathrm{adj}})^n}\cdot n! \tag{B6}$$

So the conditional probabilities A_n for both processes coincide, i.e., the joint distribution of uniformly distributed stops within the interval (0, $l_{i\,\mathrm{adj}}$) is the same as the joint distribution of the Poisson process points (stops) if there are *n* points within the interval (0, $l_{i\,\mathrm{adj}}$). It should be noted that for the

Poisson process the random length of an interval between consecutive points is exponentially distributed (Cox & Smith, 1961). It testifies that these two approaches to study $l_{ij\,\mathrm{adj}}$ point to the exponential distribution of $l_{k\,\mathrm{adj}}$.

The exponential distribution of link lengths can be confirmed if we rank the points $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$, which correspond to the above mentioned order statistics ($x_{(0)} = 0 \le x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)} \le x_{(n+1)} = l_{ij\,\mathrm{adj}}$), and determine the distribution of the value ($x_{(k+1)} - x_{(k)}$) , where k =0, 1, 2,…, n. In order to do that, it is expedient to determine the probability

$$P\{x_{(k+1)} - x_{(k)} > x\}, \quad k = 0,1,2,\ldots,n, \tag{B7}$$

where:

$\{x_{(k+1)} - x_k > x\}$ Event that there is no stop within an interval x (see Figure B4)

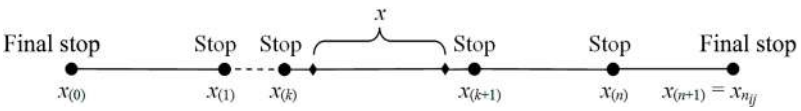


Figure B4: Graphic representation of event $\{x_{(k+1)} - x_k > x\}$ that there is no stop within an interval x

Considering Figure B4, it is possible to determine the probability of the event presented by Equation B7 as the probability of the product of independent events means that there are no points $x_{(1)}, x_{(2)}, \ldots, x_{(n+1)}$ beyond the interval with the length x (Saaty, 1961):

$$P\{x_{(k+1)} - x_{(k)} > x\} = \left(\frac{l_{ij\,\mathrm{adj}} - x}{l_{ij\,\mathrm{adj}}}\right)^n = \left(1 - \frac{x}{l_{ij\,\mathrm{adj}}}\right)^n. \tag{B8}$$

Let $l_{ij\,\mathrm{adj}} \to \infty$ and n $\to \infty$ in such a way that the limit $l_{ij\,\mathrm{adj}} / (n+1) = l_{ij\,\mathrm{adj}} / n_{ij} \to \bar{l}'_{k\,\mathrm{adj}}$ is fixed, i.e., the average link length except $l_{\min}$ tends to $l_{k\,\mathrm{adj}}$. Then we can preliminary state that

$$\lim P\{x_{(k+1)} - x_{(k)} > x\} = e^{-\frac{x}{\bar{l}_{k\,\mathrm{adj}}}}, \; x > 0. \tag{B9}$$

To prove it, let $x > 0$ and be fixed. According to the definition of the natural logarithm and taking the theorem on passage to the limit under the sign of the continuous function, we can have (Saff & Snider, 1976):

$$\lim P\{x_{(k+1)} - x_{(k)} > x\} = \lim\left(1 - \frac{x}{l_{ij\,\mathrm{adj}}}\right)^n = e^{\lim n \ln\left(1 - \frac{x}{l_{ij\,\mathrm{adj}}}\right)}. \tag{B10}$$

As it is accepted that $l_{ij\,\mathrm{adj}} \to \infty$, then $\frac{x}{l_{ij\,\mathrm{adj}}} \to 0$ and $\ln\left(1 - \frac{x}{l_{ij\,\mathrm{adj}}}\right) \sim -\frac{x}{l_{ij\,\mathrm{adj}}}$. Therefore,

$$\lim P\{x_{(k+1)} - x_{(k)} > x\} = e^{\lim n \cdot \ln\left(1 - \frac{x}{l_{ij\,\mathrm{adj}}}\right)} = e^{\lim\left(-\frac{n \cdot x}{l_{ij\,\mathrm{adj}}}\right)} = e^{-x \cdot \lim \frac{n}{l_{ij\,\mathrm{adj}}}} = e^{-x \lim \frac{n+1}{l_{ij\,\mathrm{adj}}} \cdot \lim \frac{n}{n+1}} = e^{-x \cdot \lim \frac{n+1}{l_{ij\,\mathrm{adj}}}} = e^{-\frac{x}{\bar{l}_{k\,\mathrm{adj}}}}. \tag{B11}$$

This confirms the correctness of Equation B9.

Thus, the hypothesis about the exponential appearing of some intermediate stops between an existing final stop in the PT system and a new final stop in the expanded city area is valid.

Appendix C: Determination of the link length distribution when the links are the components of the distances between pairs of stops in a city transit system

We can designate a random variable—the length of the *s*-th link on the route between stops *i* and *j*—by the symbol $l_{s\,\mathrm{adj}}$. When the distribution law of the variable $l_{k\,\mathrm{adj}}$ is known, the probability of that the $l_{s\,\mathrm{adj}}$ lies within an interval $(a;b]$ is determined by the equation

$$P\{a < l_{s\,\mathrm{adj}} \le b\} = F(b) - F(a) = \int_a^b f(l_{k\,\mathrm{adj}})dl_{k\,\mathrm{adj}}, \tag{C1}$$

where:

$F(a),F(b)$	Values of the distribution function of the variable $l_{k\,\mathrm{adj}}$ at the points a and b respectively
$f(l_{k\,\mathrm{adj}})$	Density function of the variable $l_{k\,\mathrm{adj}}$

Let $p = P\{a < l_{s\,\mathrm{adj}} \le b\}$, $q = P\{l_{s\,\mathrm{adj}} \notin (a;b]\}$ and the distance $l_{ij\,\mathrm{adj}}$ has the n_{ij} links. Let us introduce $v_{(a;b]}$, which is a random variable denoting the number of links within an interval $(a;b]$. Supposing the link lengths are independent and identically distributed we can determine the probability of the event that exactly m links along the distance that consists of n_{ij} links, are within an interval $(a;b]$:

$$P\{\nu_{(a;b]} = m\} = C_{n_{ij}}^{m} \cdot p^{m} \cdot q^{n_{ij}-m}, \quad m = 0,1,2,\ldots,n_{ij}. \tag{C2}$$

On the other hand, $v_{(a;b]}$ can be presented as the sum of indicators

$$I_1^{(a;b]} + I_2^{(a;b]} + \ldots + I_{n_{ij}}^{(a;b]} = \nu_{(a;b]}, \tag{C3}$$

where:

$$I_s^{(a;b]} = \begin{cases} 1, \text{ if } l_{s\,\mathrm{adj}} \in (a;b] \text{ with probability } p_1; \\ 0, \text{ if } l_{s\,\mathrm{adj}} \notin (a;b] \text{ with probability } q_1 \end{cases} \tag{C4}$$

is an indicator of a random event {the length of the *s*-th link is within an interval $(a;b]$}.

The indicator of a random event $\{l_{s\,\mathrm{adj}} \in (c;d]\}$ for another interval $(c;d]$ can be determined in an analogous manner:

$$I_s^{(c;d]} = \begin{cases} 1, \text{ if } l_{s\,\mathrm{adj}} \in (c;d] \text{ with probability } p_2; \\ 0, \text{ if } l_{s\,\mathrm{adj}} \notin (c;d] \text{ with probability } q_2. \end{cases} \tag{C5}$$

Let $p_1 \ge p_2$, i.e., the probability that a link length is within an interval $(a;b]$ is more than or equal to the probability that a link length is within an interval $(c;d]$. If we introduce an elementary consequence of a random experiment that a certain point is within an interval $(0;1]$ — $\omega_s \in [0;1]$, the above-mentioned indicators can be compared within one probability space

$$I_s^{(a;b]}(\omega_s) \ge I_s^{(c;d]}(\omega_s) \tag{C6}$$

and they can be represented by the Bernoulli distribution (Forbes et al., 2011) that is shown in Figure C1.

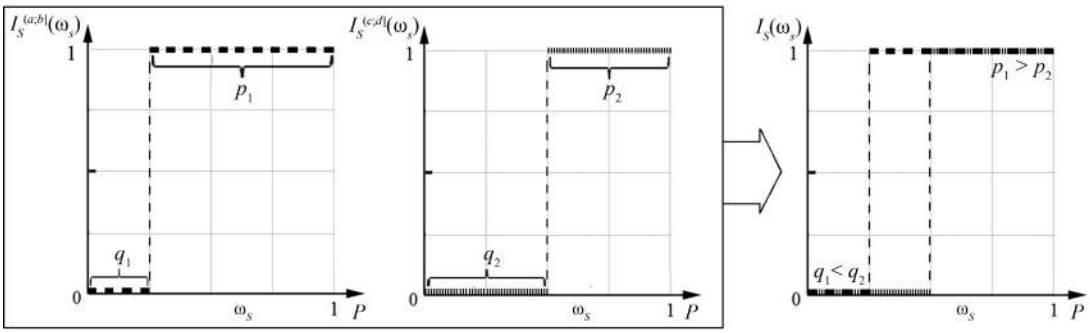


Figure C1: Distribution of the indicators of events

Let $\omega = (\omega_1, \omega_2, ..., \omega_{n_{ij}})$ be an elementary consequence of the random variables $v_{(a;b]}$ and $v_{(c;d]}$. Then

$$I_s^{(a;b]}(\omega) \equiv I_s^{(a;b]}(\omega_s), \ s = 1,2,...,n_{ij} \tag{C7}$$

and

$$v_{(a;b]}(\omega) = I_1^{(a;b]}(\omega) + I_2^{(a;b]}(\omega) + ... + I_{n_{ij}}^{(a;b]}(\omega). \tag{C8}$$

Now, it is necessary to verify the following statement: if

$$P\{a < l_{s\,\text{adj}} \le b\} = p_1 \ \ge \ P\{c < l_{s\,\text{adj}} \le d\} = p_2, \tag{C9}$$

then

$$P\{v_{(a;b]} \ge m\} \ge P\{v_{(c;d]} \ge m\}, \ m = 0,1,2,...,n_{ij}, \tag{C10}$$

i.e.,

$$v_{(a;b]} \underset{P}{\ge} v_{(c;d]}. \tag{C11}$$

Under the condition $p_1 \ge p_2$ and $(q_1=1-p_1) \le (q_2=1-p_2)$ when $p_1+q_1=1$ and $p_2+q_2=1$, it follows that

$$I_s^{(a;b]}(\omega) \ge I_s^{(c;d]}(\omega), \ s = 1,2,...,n_{ij} \tag{C12}$$

Therefore,

$$v_{(a;b]}(\omega) = I_1^{(a;b]}(\omega) + I_2^{(a;b]}(\omega) + ... + I_{n_{ij}}^{(a;b]}(\omega) \ge \tag{C13}$$
$$\ge v_{(c;d]}(\omega) = I_1^{(c;d]}(\omega) + I_2^{(c;d]}(\omega) + ... + I_{n_{ij}}^{(c;d]}(\omega).$$

If we compare these values within the same probability space $\omega = (\omega_1, \omega_2, ..., \omega_{n_{ij}})$, it is clear that

$$v_{(a;b]}(\omega) \ge v_{(c;d]}(\omega) \tag{C14}$$

This testifies that the statements made by means of Equation C9, Equation C10, and Equation C11 are correct.

It means that the higher (according to the distribution law) the probability of the appearance of links with certain length $l_{k\,\mathrm{adj}}$, the more frequently such links will become the components of the distances between pairs of stops when forming the variable $l_{s\,\mathrm{adj}}$. This result confirms the hypothesis, which is made in the second paragraph of the third subsection of the Fundamentals, and can be verified using actual data.

Appendix D: City characteristics

Table D1: Population and area of the cities

City	Index	
	Population	Area, km^2
Kyiv	2,868,300	847.7
Kharkiv	1,445,200	350.0
Kryvyi Rih	652,000	415.0
Sumy	270,000	145.0
Kirovohrad	232,000	103.0
Sverdlovsk	98,000	83.8
Oleksandrija	83,000	55.0
Izium	50,900	41.0
Balaklia	29,400	162.7
Kupiansk	25,100	33.4

Note. Adapted from Public Information (http://land.gov.ua/), by State Service of Ukraine for Geodesy, Cartography & Cadastre, 2017. Copyright 2017 by the State Service of Ukraine for Geodesy, Cartography & Cadastre; Statistical Information (http://www.ukrstat.gov.ua/), by State Statistics Service of Ukraine, 2017. Copyright 2017 by the State Service of Ukraine for Geodesy, Cartography & Cadastre.

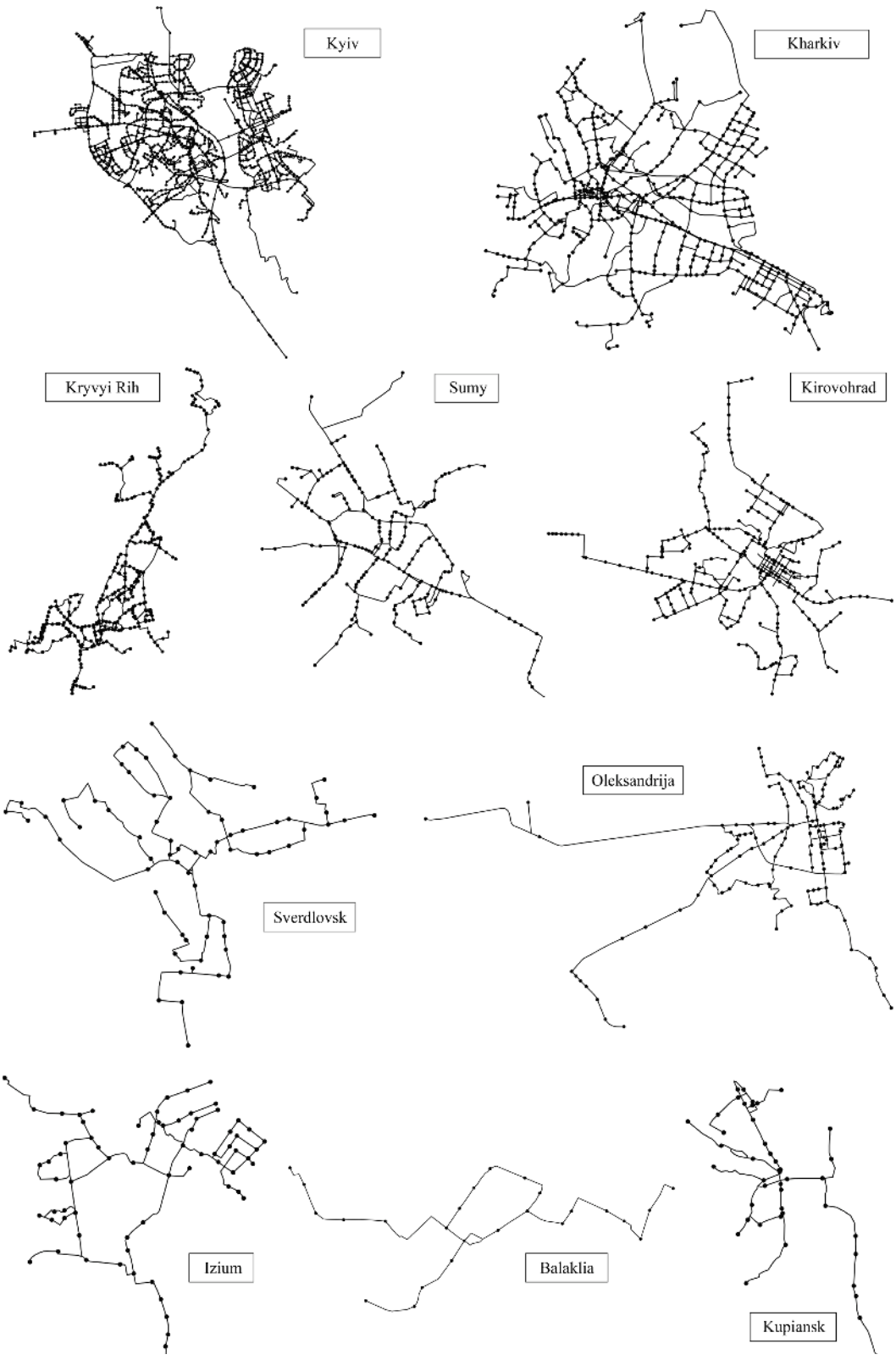


Figure D1: Topology of transit networks in 10 Ukrainian cities