Transport and Telecommunication, 2010, Volume 11, No 2, 4–11
Transport and Telecommunication Institute, Lomonosova 1, Riga, LV-1019, Latvia

CONSTRUCTION OF THE URBAN PUBLIC TRANSPORT SYSTEM'S QUALITY INDICATOR WITH MISSING DATA

Irina Yatskiv[1], Irina Pticina[2]

Transport and Telecommunication Institute
Lomonosova 1, Riga, LV-1019, Latvia
[1]Ph.: +371 67100544. E-mail: ivl@tsi.lv
[2]Ph.: +371 67100596. E-mail: jurina@tsi.lv

The purpose of this investigation is the development of transport system's quality indicator – UPTQI (urban public transport quality indicator) calculation methods where some values are missing in the data set. As a quality indicator, the indicator accumulating the information about urban public transport system – composite indicator is used. The algorithm described in [1] has been used for developing the composite indicator. The special accent is made on a choice of a missing data imputation method in an initial data set and on stability of results. It is not possible to make the unequivocal recommendation for use concrete method in the presence of a large number of offered methods. In many respects it depends on the initial data. The initial data concerning the cities of Europe were taken from EUROSTAT database [2]. The 8 sub-indicators for 2006 have been used for constructing the UPTQI.

The research was divided into three stages. At first, artificial missing data was introduced, and investigated the influence of the 3 selected methods of missed data substitution (unconditional mean imputation, imputation by median and clustering-based imputation) upon the results of composite indicator calculation. In the next stage the cities with the missing data have been added (in total 62 European cities). The method which is chosen as the best for this problem during the first stage has been used to the missing data imputation. To calculate the weights and to aggregate the same variants of algorithm, as at the first stage were used. At the last the results of UPTQI constructing were analysed for uniformity.

Keywords: public transport, quality, composite indicator, weights, missing data, imputation

1. Introduction

The purpose of public transport is rendering safe, reliable, punctual, accessible, non-polluting and cost-effective transport services to people. So, there are many characteristics, which need to be measured for understanding the actual quality of public transport system, and it is difficult to estimate the actual situation on the basis of this multivariate set of data. The benchmarking of the urban public transport can apply a scalar integral or composite indicator to be constructed – instead of using a set of initial characteristics as a basis. Indicator is an integral performance index presenting a complex estimate of a process, system, or object. A multivariate set of sub-indicators forms a basis of developing the integral indicator which is transformed into a scalar in a certain way.

The integral (composite) indicator is a function from sub-indicators and weights as follows:

$$CI_i^t = f(x_{i,1}^t, x_{i,2}^t, ...x_{i,m}^t, w_1, w_2, ..., w_m), \quad (1)$$

where CI_i^t – value of indicator for object i ($i = 1...n$) at time t,

$x_{i,j}^t$ – value of sub-indicator j ($j = 1...m$) for object i at time t,

w_j – weight associated with sub-indicator j ($j = 1...m$).

The most important advantages of composite indicator in the scalar form are a possibility of using it successfully for comparison and development analysis instead of a multivariate set of parameters. The European Plan of Research in Official Statistics (EPROS) for 2007–2013 singles out the continuation of work in the field of composite indicators and applying statistical methods in developing them as one of the top priorities.

There is the attempt to construct a composite indicator of urban public transport quality indicator – UPTQI, provided for by urban public transport system (UPTS) in the work.

One of the essential problems arising when UPTS quality is assessed is the missing data problem. The approaches to missing values imputation can be subdivided into two groups: Single imputation (Implicit modelling – Unconditional mean/median/mode imputation, Regression imputation, Expectation Maximisation imputation; Explicit modelling) and Multiple imputation. As there are no universal recommendations for usage of this or that method of the imputation of the missing data and results of its usage depend on character of a solved problem, on a set of variables, and on model of skips. In order to choose

the imputation method, first of all we will conduct the research by definition of the best method for our set of objects and variables.

The purpose of this research is the methodology development of estimation the urban public transport system's quality indicator in case where some values of initial characteristics are missing.

2. Research Methodology

Let's assume that there are values of *M* sub-indicators describing urban system of public transport for *L* cities. A part of cities (*N*) do not have any skips in sub-indicator values, while (*L-N)* cities do have gaps in data.

The investigation will be performed in three stages. The research algorithm is presented on Figure 1.

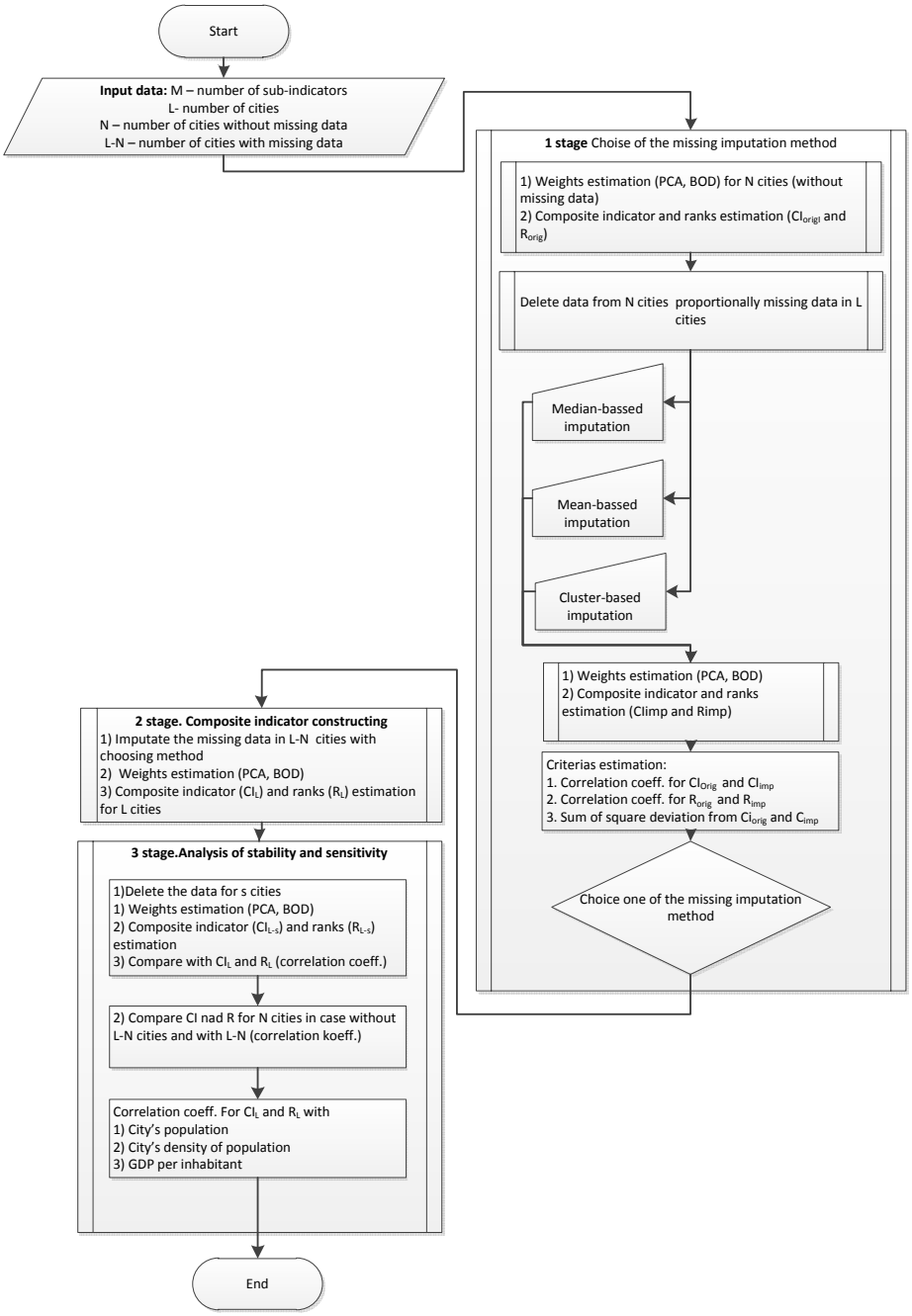

Figure 1. The research algorithm

At the first stage, we shall investigate the influence of missing data imputation methods on the composite indicator value. Let's remove some values from the data set where sub-indicator values are known (N cities) – pro data missing values in the complete data set (L-cities). For imputation of the missing data in considered data set the following methods were used:

- unconditional mean imputation;
- imputation by median;
- clustering-based imputation.

The method of *unconditional mean imputation* is the simplest one. It has been included into the investigation as the method most frequently used in statistical software. It implies estimation of missing values $x_{i,j}$ by the average value $\overline{x_j}$.

$$\overline{x_j} = \frac{1}{n_j}\sum_{i=1}^{n_j} x_{i,j} \,, \tag{2}$$

where $x_{i,j}$ – value of sub-indicator j for object i,
n_j – count of objects with fully observed sub-indicator j.

The *median* of the distribution could be calculated on the available sample and to substitute missing values.

Various approaches for missed data imputation, implying cluster analysis, are known. In this work, we consider the *clustering-base missing data imputation* based on the following steps:

1. One of the methods of cluster analysis is applied to objects without missing data, – and C clusters are singled out.
2. The distance to the centres of all the C clusters is calculated with respect to each i object of observation having some missing data:

$$D_{i,c} = \sqrt{\sum_{j=1}^{m}(x_{i,j} - \overline{x_{j,c}})^2} \,, \tag{3}$$

where $D_{i,c}$ – distance between object i and cluster c,
$x_{i,j}$ – value of sub-indicator j (no missing) of object i,
$\overline{x_{j,c}}$ – mean value of sub-indicator j of c-cluster's objects.

Any variables where values are missing are not involved in the calculation of distance.

3. The nearest cluster is determined, with the minimal distance to it.
4. The skipped value of i object is substituted for the mean value of the corresponding variable pertaining to those observations that are attributed to the nearest cluster.

We shall calculate the value of composite indicator (CI) (Eq. 1) and rankings (R) of N cities with respect to 3 cases of missing values imputation. For constructing the composite indicator it was used the ten-step algorithm, which was developed the Organization of Economic Cooperation and Development (OECD) [1]. For calculating weights and aggregating primary indices into the composite indicator methods based on Equal weighting approach, Principal Components and Factor analysis (PCA/FA) model and benefit of the doubt approach (BOD) were considered [1]. Afterwards, the received CI and R values will be compared to original values CI_{orig} and city rankings R_{orig} that had been calculated for the same cities – however, with no skip in data. Then the most appropriate method of missing data imputation will be selected for the given set of data. We shall use as selection criteria:

- the Person's correlation between values CI_{orig} and CI, given in these stage;
- the Spearman's correlation between values R_{orig} and R, given in these stage;
- the sum of deviation square CI from CI_{orig} .

At the second stage, we shall impute the missed values in the full set of data (L cities) by using the method selected at the first stage. We shall calculate weight coefficients, develop composite indicators for all the cities (CI_L) and determine their rankings (R_L).

At the third stage, we shall perform the analysis for sensitivity and stability of results. To do that, we shall calculate the composite indicator value by deleting some cities from the full list, – and then compare the obtained values with R_L and CI_L. Moreover, we shall compare values CI and R that had no skipped data (N cities) – both when data on other cities is available and missing.

3. Construction of Urban Public Transport Quality Indicator

3.1. The Initial Data

Let us consider as the object – the urban system of public transport, as composite indicator – urban public transport quality indicator – UPTQI and as the sub-indicators – the particular quality characteristics. The database EUROSTAT has been analysed about presence of the UPTS data for the European cities. The data from national sources has not been used, as they are not always calculated with use of one methodology. 21 indicators, which characterise the urban public transport system, have been found in data base EUROSTAT [2]. Unfortunately, the values of mentioned indicators in the database are missing for many cities. Therefore, in the given research the group of the cities (objects) with the most set of the indicators values has been used and special attention in the given work was attend to the quality of the composite indicator construction in case with the missing data. To conduct the investigation, the data describing UPTS in 62 European cities were used with respect to 2003–2006 moments of time. The values of 8 indices shown in Table 1 were used in this investigation.

Table 1. List of sub-indicators

№	Sub-indicators	Code
1	Proportion of journeys to work by public transport (rail, metro, bus, tram)	x1
2	Length of public transport network / land area	x2
3	Number of stops of public transport per km2	x3
4	Cost of a monthly ticket for public transport (for 5–10 km)	x4
5	Number of stops of public transport per 1000 pop.	x5
6	Number of stops per 1 km of public transport network	x6
7	Proportion of public transport network on fixed infrastructure /Proportion of public transport network on flexible routes	x7
8	Proportion of the area used for transport (road, rail, air, ports)	x8

Sub-indicators *x2, x3, x5, x6* and *x8* characterize the availability of UPTS, *x4* – the economic component; *x7* – the environmental impact (since the fixed-route transport uses alternative fuel (as electricity for example). *x1* – may be attributed both to accessibility indicators and subjective indicators – i.e., the attitude towards the quality service as offered by the system: tenants tend to prefer the system more frequently if it provides a higher quality. Unfortunately, no data on other characteristics could be found – like, for example, indices that would characterize the safety of UPTS in these cities.

To construct the indicator with respect to the complete set of data, we shall use the data describing 62 European cities, from them: 37 German cities (without missing data) and 25 other European cities (basically capitals, 3 from them don't have missing data and 22 with missing data in sub-indicators values, but no more than 3 missing values). The largest number of skips in the complete information is observed in values of following sub-indicators: *x1* – Proportion of journeys to work by public transport (rail, metro, bus, tram) and *x8* – Proportion of the area used for transport (road, rail, air, ports)). The total number of missed values is 34 out of 496 which makes 6.85% from the total number of values. Out of that number, 17 values are missing with respect to variable *x1* (27% from 62 values), and 8 values with respect to variable *x8* (13% from 62 values.).

3.2. Analysis of the Influence of Missing Data Imputation Method on Composite Indicators' Scores

In the previous investigation [3] the composite indicator describing the quality of urban public transport was developed for 37 cities of Germany having no gaps in data. We shall call these values original CI_{orig} and R_{orig}.

To investigate the influence of missing data imputation methods on the composite indicator value, let's delete approximately 6.85% of the values from the data describing 37 German cities pro rata the volume of missing values in the complete initial data set; out of them, 10 values with respect to variable *x1*, 5 values with respect to variable *x8*, and 5 random values with respect to other variables. Let us substitute the missing values by using 3 methods of missing data imputation: unconditional mean imputation, imputation by median and clustering-based imputation. Let's calculate the composite indicator for all the three cases. We shall use two methods – PCA and BOD to calculate the weighting factor.

Figures 2 and 3 show normalized values of the composite indicator (CI_{PCA} and $CI_{\boldsymbol{BOD}}$) calculated with respect to cases implying substitution of missed values by using three methods – for 4 cities. The obtained values are presented as compared to original values.

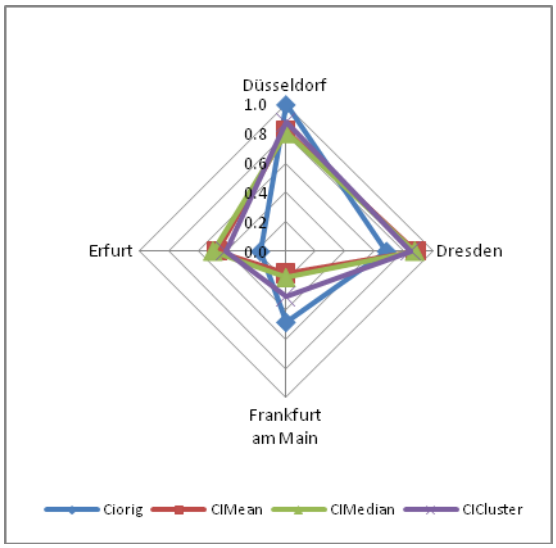

Figure 2. The values of $CI_{PCA}norm$

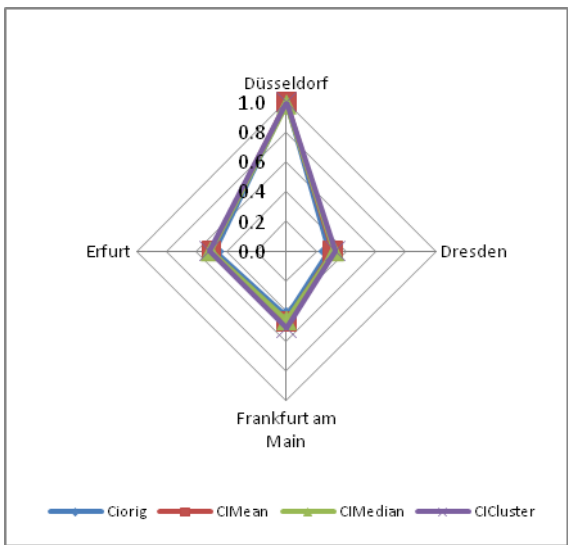

Figure 3. The values of $CI_{BOD}norm$

Table 2 and Table 3 shows the correlation ratio between the original values CI_{orig} and R_{orig} with the values received in the case implying gaps in data, where the weighting factors are calculated by two methods – BOD and PCA. Table 4 shows the values of the sum of squared deviations from the true value of composite indicator.

Table 2. The value of Pearson's correlation with originals CI norm

Methods	PCA	BOD
Mean	0.867	**0.971**
Median	0.867	0.967
Cluster	0.914	**0.971**

Table 3. The value of Spearman's correlation with originals R

Methods	PCA	BOD
Mean	0.815	0.941
Median	0.806	0.932
Cluster	0.837	**0.950**

Table 4. The sum of squared deviations for *CIorig* and *CI* what calculated with imputation

Methods	Mean	Median	Cluster
PCA	0.0180	0.0175	0.0120
BOD	**0.0047**	0.0054	0.0052

The largest value of correlation coefficient between the calculated *CI* and CI_{orig} , also, between *R* and R_{orig} is observed for values obtained by the BOD method (the values are marked in Tables 2 and 3). Besides, applying the method of missing data imputation by mean data yields the composite indicator value closer to the original one; however, one yields the rank city value closer to the original when missing data is replaced by cluster-based missing data imputation approach. Moreover, the BOD method yields the least value of SSD (sum of squared deviations) when substituted by mean values – 0.0047.

So, to calculate the weighting coefficients, we shall use the BOD approach as the least sensitive to gaps in data.

3.4. The Estimation of the Composite Indicator Values for 62 European Cities

Let's calculate the composite indicator for the full data set – 62 cities. Then, we impute the missing values by using the method based on the cluster analysis. As a result of applying the cluster analysis, 4 clusters have been singled out for 37 German cities. Let's determine the cluster to which the distance is the least one, for each of 22 cities having gaps in data. Then we substitute the missing values by the average value of missing variable in the cluster selected. Further, we calculate the weighting coefficient w_i and the composite indicator value CI_{62} by using BOD approach. Table 5 and Figure 3 show the values of CI_{62}, the normalized values of composite indicator ($CI_{62}norm$), ad the rankings (R_{62}) for some cities. The cities have been ranked in descending order of the normalised CI_{62} – from Torino ($CI_{62}norm = 1$) to Schwerin ($CI_{62}norm = 0$).

Table 5. The values of composite indicator and ranks of cities

Ranks	Cities	CI_{62}	CI_{62} norm
1	Torino	1.018	1
2	Madrid	1.013	0.992
3	Helsinki	0.979	0.940
4	Bratislava	0.945	0.889
5	Brussel	0.935	0.874
6	Rotterdam	0.855	0.753
7	Warszawa	0.842	0.733
8	Kosice	0.827	0.711
9	Genève	0.819	0.699
10	Wien	0.819	0.698
…			
18	Berlin	0.692	0.506
…			
22	Dresden	0.664	0.463
23	Düsseldorf	0.654	0.449
…			
26	Riga	0.625	0.406
…			
61	Bielefeld	0.362	0.008
62	Schwerin	0.357	0

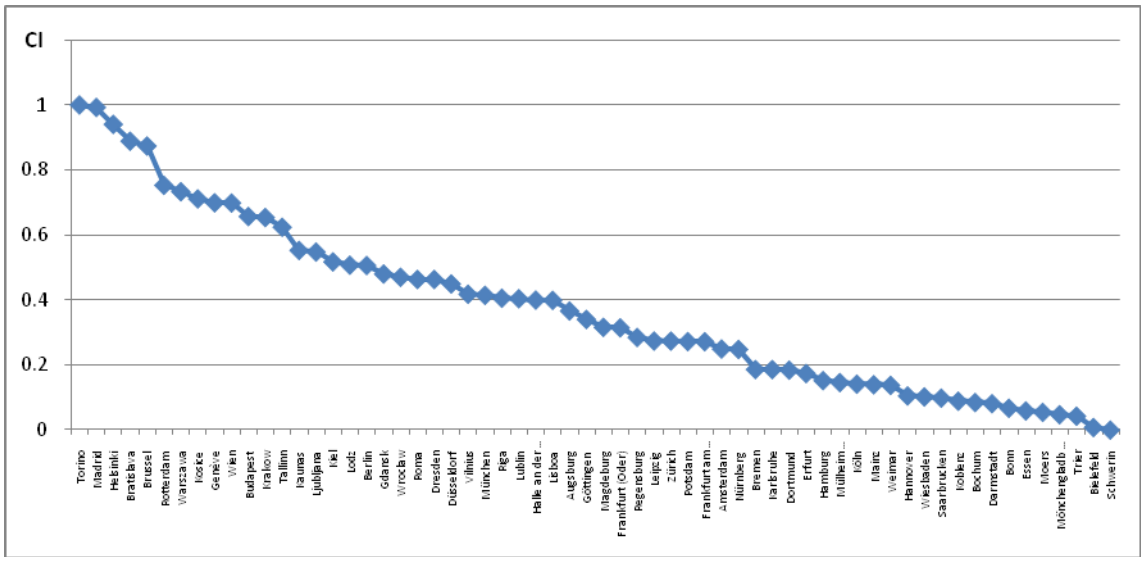

Figure 4. The values of CInorm for 62 European cities

3.5. The Sensitive and Stability Analysis

To perform the sensitivity analysis of the obtained results, the data describing 3 cities was deleted from the cities lit (one by one):

- Torino (ranking first in the ratings of cities);
- Lisbon (the largest number of gaps in data – 3 missing values);
- Ljubljana (the indeterminate value in the variable – Proportion of public transport network on fixed infrastructure/Proportion of public transport network on flexible routes).

For each of 3 cases, composite indicator values were calculated and compared to the value CI_{62}. High values of correlation coefficient (the value of Pearson's correlation > 0.98) attest to the stability of the yielded results.

Also, for stability analysis, let's compare *CI* values to the rankings of 37 German cities yielded at the first and the second stages. Figure 5 shows the rankings of 37 German cities calculated both with and without using any data describing other cities (R_{37} and R_{62}). In both cases, BOD method was used to calculate the weights' values.

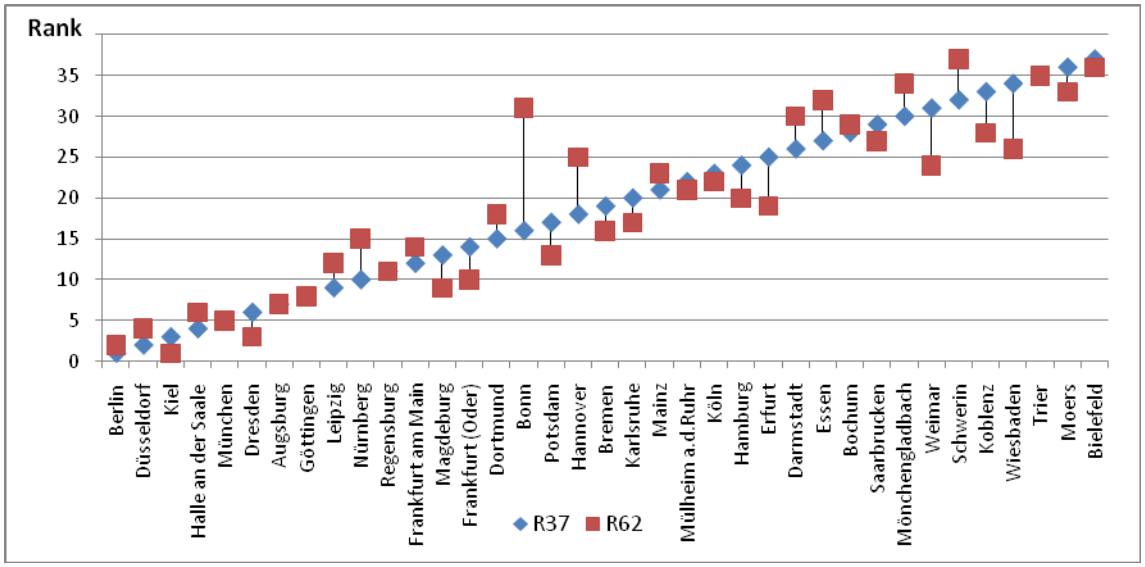

Figure 5. The ranks of 37 German cities

The high value of the correlation ratio between normalised *CI* values (the value of Pearson's correlation 0.94) and the rankings yielded for 37 German cities – both with and without other European cities included (the value of Spearman's correlation – 0.92) attest to the stability of the yielded results.

4. Conclusions and Discussion

In the course of the investigation, a composite indicator for 62 European cities with missing data was developed. The clustering-based method was selected as the most appropriate one for imputation of missing data. For weights calculation the BOD method was used as the least sensitive to gaps with respect to the given set of data.

Moreover, a methodology of composite indicator calculation in case with missing data was developed. The methodology includes the steps as follows:

1. Calculation of composite indicator for cities without missing data in sub-indicator values.
2. Simulation of gaps within the data set with no gaps pro rata the number and the distribution between variables in the full set of data.
3. Substitute of missing values by using a few methods and calculation of composite indicator values for each specific case.
4. Selecting the most appropriate method of missing data imputation by using the follows criteria:

- the Person's and Spearman's correlation coefficients with original *CI*;
- the Spearman's correlation coefficients with original ranks of cities;
- the sum of deviation square from original *CI*.

5. Substitute the missing values in the full data set by using the method selected at the previous stage, and calculate the composite indicator value with respect to the full data set.
6. Perform sensitivity and stability analysis of the yielded results.

The obtained results stated in the article provide a basis for the further investigation of composite indicator construction methods to be used for estimating the Public Transport System.

Acknowledgments

The article is written with the financial assistance of European Social Fund. Project No 2009/0159/1DP/1.1.2.1.2/09/IPIA/VIAA/006 (The Support in Realisation of the Doctoral Programme "Telematics and Logistics" of the Transport and Telecommunication Institute).

References

1. Nardo, M. a.o. Handbook on Constructing Composite Indicators: Methodology and User Guide, *OECD Statistics Working Paper*, No 3, 2005. 108 p.
2. Statistical Office of the European Communities – EUROSTAT – http://epp.eurostat.ec.europa.eu
3. Pticina, I., Yatskiv, I. Constructing the urban public transport system quality indicator. In: *The 1st International Conference on Road and Rail Infrastructure (CETRA 2010), May 17–18, 2010, Opatija, Croatia,* pp. 223–229.