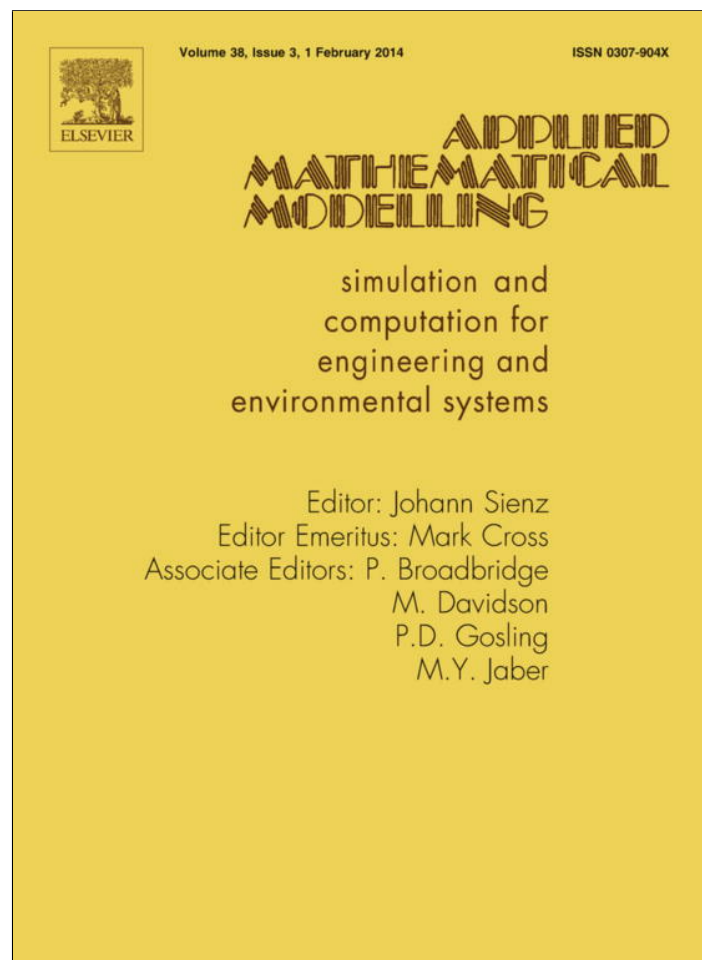




This article appeared in a journal published by Elsevier. The attached copy is furnished to the author for internal non-commercial research and education use, including for instruction at the authors institution and sharing with colleagues.

Other uses, including reproduction and distribution, or selling or licensing copies, or posting to personal, institutional or third party websites are prohibited.

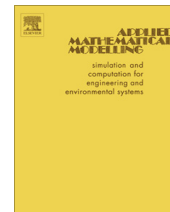
In most cases authors are permitted to post their version of the article (e.g. in Word or Tex form) to their personal website or institutional repository. Authors requiring further information regarding Elsevier's archiving and manuscript policies are encouraged to visit:

<http://www.elsevier.com/authorsrights>



Contents lists available at ScienceDirect

Applied Mathematical Modelling

journal homepage: www.elsevier.com/locate/apm

Adjustment boarding and alighting passengers on a bus transit line using qualitative information

Juan de Oña^{a,*}, Penélope Gómez^b, Enrique Mérida-Casermeiro^c^a TRYSE Research Group, Department of Civil Engineering, University of Granada, ETSI Caminos, Canales y Puertos, c/Severo Ochoa, s/n, 18071 Granada, Spain^b Department of Civil Engineering, University of Granada, ETSI Caminos, Canales y Puertos, c/Severo Ochoa, s/n, 18071 Granada, Spain^c Applied Mathematics Department, University of Malaga, E.T.S. Computer Science Engineering, Boulevard Louis Pasteur, s/n, 29071 Málaga, Spain

ARTICLE INFO

Article history:

Received 17 July 2012

Received in revised form 19 June 2013

Accepted 30 July 2013

Available online 20 August 2013

Keywords:

Fuzzy optimization

Transport planning

Public transport

ABSTRACT

Obtaining data to use in an urban public transport operation planning and analysis is problematic, particularly in urban bus transit lines. In an urban environment and for bus services, most ticketing methods can be used to record passengers getting on board but not getting off, and current methods are unable to make a proper adjustment of boardings and alightings based on the available data unless they do alighting counts. This paper presents a method whereby counts are made at fewer stops and qualitative information on alightings and/or vehicle loads between consecutive stops is used to make the boarding and alighting adjustment as a previous step to obtain the real origin and destination (O/D) of passengers allowing the O/D matrix calibration by using the loads between stops. Qualitative information can be obtained by the vehicle's driver or an on board observer, avoiding the necessity of counting many stops in planning period. The method is applied to a real bus transit line in Malaga (Spain) and to a set of 50 different bus transit lines with number of stops ranging from 10 to 75. The results show that the proposed method reduces the adjustment errors with regard to traditional methods, such as Least Square Method, even in the situation where no qualitative information is used. When qualitative data is used on alightings and loadings, the reduction of the average error is over 50%.

© 2013 Elsevier Inc. All rights reserved.

1. Introduction

When planning public transport networks, it is crucial to know the real origin and destination (O/D) of passengers. Surveys about the O/D of travelers are mandatory to obtain this information at every transport system. Once the O/D matrix has been obtained (based on the survey), it has to be calibrated with collected data. For that aim, in the case of bus services, the number of passengers between the bus stops (bus loads) is key information. To get this information, the transport planner needs to know the actual in and out movements of passengers at each stop along the line. Besides, bus loads are also crucial in the service operation activities, such as when deciding if an additional vehicle is required because the maximum load has been overtaken at peak time, helping to adapt the service to the demand as much as possible. Regarding to urban transit buses, collecting data on passenger boardings has progressed with the new electronic ticketing systems, like the smart card as a payment option as can be seen in the literature review made by Pelletier et al. [1], and thanks to the increased sophistication of mobile communication technologies [2]. Smart cards improve the quality of data [3] and the ticket validation systems provide information on the number of boardings. Therefore, this information is quite accurate and the only errors are due to potential device failures.

* Corresponding author. Tel.: +34 958 24 99 79; fax: +34 958 24 61 38.

E-mail addresses: jdona@ugr.es (J. de Oña), penelopegi@ugr.es (P. Gómez), merida@ctima.uma.es (E. Mérida-Casermeiro).

However, the systems cannot be used to obtain data on the number of alightings, so passenger detection systems and surveys on board or at the stops are needed for that purpose. Several surveyors may be needed if there are several exits (e.g., in articulated buses) and high passenger volumes. Such data collection is much more costly and subject to more errors than boarding counts. So, improved techniques for collecting data on transit operation are essential to improvements in transit operating efficiency. Two-time mode cards have been adopted in certain exceptional cases [4] (i.e., Beijing Municipal Government Public Traffic) to record where passengers board and alight. Card scanners are placed at the entrance and at the exit, but the systems are not used on most of transport services at a global level, passenger tickets need to be scanned twice which means double investment.

New emerging technologies are being developed, such as images recognition, weight sensors or counting sensors but, so far, the pilot project experiences have failed because they still present too many errors (i.e., open field, shadows, partial vision, etc.) and it seems to give erroneous information, which at the end must be used as fuzzy data, that no traditional method is able to work with.

It is important to remember that in both, the case of interurban and underground transport systems, where passengers buy the ticket before boarding and in many underground networks the passengers must scan their tickets before they exit, this method would be useless. But it still remains a wide field to be applied on urban or metropolitan bus lines worldwide.

2. Background

Several methods have been developed to adjust data on a transit line when both boarding and alighting data are available [5]. In general, all methods seek to narrow the gap between observed values and adjusted values as much as possible, subject to contour conditions.

The existing methods can be classified into two groups, depending on the nature of the observed values and how they are processed:

- Group One: The adjusted values are based on their closeness to the observed values. The methods used are: the least squares method (LSM); the maximum likelihood adjustment; and the fuzzy regression adjustment [6]. In addition to the above methods, other authors have defined a stochastic method in which it is assumed that passenger boardings follow a Poisson distribution and the number of passengers alighting follows a binomial distribution [7].
- Group Two: These methods assume that the observed value is approximate and that the adjusted value is within a range created around the observed values. This group can include fuzzy optimization and the required interval regression adjustment. Using the fuzzy sets theory, fuzzy optimization adjustments allow soft constraints to be added to the relationships between volumes at transport nodes, seeking data reliability and the relationships between volumes. The adjustment with the required interval regression seeks the adjusted value within a crisp contour. This method is appropriate for those cases in which the analyst does not trust the accuracy of the observed data.

All the above methods require quantitative data to be able to make the adjustment, and obtaining such data is expensive. On the other hand, information on vehicle loads between stops is not often used to make the adjustment between boarding and alighting data. Rather, it is the final output of the adjustment.

At almost no extra effort, qualitative information on the number of passengers who alight at a stop or on loads between stops on a transit line could be obtained, along the lines such as: a few passengers, many passengers, half the load, or I do not know how many alighted at stop x_i ; the bus was half full, almost empty or half full between stop x_i and x_{i+1} .

The above-mentioned methods are not able to use qualitative information, however. Although the methods in Group Two use fuzzy logic, they are based on quantitative values, so they can only be applied if a quantitative value is assigned to each observed value. Doing so would add an element of randomness to the results obtained. To explain this, let us suppose that there are five stops on a line and the boarding data is available (80, 20, 20, 20, 0) but the number of passengers alighting could not be quantified. To be able to apply the existing methods, a quantitative value would need to be assigned to each alighting. If that information is not available, one analyst could suppose that (0, 0, 0, 0, 140) have alighted, whereas another analyst might suppose (0, 80, 20, 20, 20). The results obtained by both analysts would be completely different.

In this paper, we present a new method that uses fuzzy optimization based on qualitative information about the number of passengers alighting at each stop and about the vehicle load between stops. The aim of the method is to use this information to enhance boarding and alighting adjustments, with two possibilities:

- One, the information on the alightings provided by surveyors (quantitative, at a high effort and cost) could be replaced by qualitative information on the number of passengers who alight at each stop and on the vehicle load between stops, which could be provided by the vehicle's driver. This would dispense with the need to hire surveyors to do the job, with the resulting financial saving.
- Two, to see the percentage of alightings that would not need to be counted while retaining the adjustment's accuracy, if we used qualitative information on the vehicle load between stops provided by the vehicle's driver.

This paper is organized as follows: Section 2 describes the method and the computational issues; in Section 3 the method is applied to a real transit line and, in order to verify the results, it is applied to a set of different types of lines; Section 4 discuss results; and, finally, Section 5 presents the main conclusions of the paper.

3. Theoretical approach

3.1. Description of the problem

Given a transit line with N stops, we want to adjust passenger boardings and alightings at each stop, as well as the loads between two consecutive stops, based on information obtained by several different methods, in such a way that the following basic principles of flow conservation are met:

- the total number of boarding passengers should be equal to the total number of alighting passengers, and
- the number of passengers on board between stops k and $k + 1$ should be greater than zero and less than the vehicle capacity ($L_{\max}$)

The initial variables and data for solving the problem are the values for passenger boardings and alightings, vehicle loads between stops and $L_{\max}$.

The data collection can provide several types of information: quantitative numerical data (precise integer values or with an error), qualitative data (many, a few, etc.) or missing data (no information is available on the value adopted by a specific variable).

The type of data will depend on the variable taken into consideration:

- Passenger boardings: these are obtained by the ticket sales method, so it can be assumed that there are no errors and therefore the values are deemed to be exact fixed integers. In a context of scarce information, the few data with small or null error (only in the case of potential failures in the devices) will be considered as fixed data.
- Passenger alightings: depending on the method used for data collection, it can be quantitative numerical data with errors (from counts), qualitative information (from the perception of an analyst or driver), or missing information (when no information is available).
- Vehicle load between stops: it may be considered as qualitative information (from the perception of an analyst or driver) or missing (if an analyst or driver has not additional information on loads).
- Capacity of the vehicle ($L_{\max}$): it is considered to be a fixed numerical value, used as a framework for establishing the different categories of qualitative information (many, some, few, etc.).

3.2. The proposed method

The first step to solve this problem is to use membership functions to represent the above concepts. Fig. 1 shows the membership functions for four concepts: fixed number (quantitative information with no error); crisp number (quantitative information with errors); fuzzy information (qualitative information) and; missing value.

A membership function is convenient for representing the idea that the adjusted value should be “close” to the observed value and the acceptability of the adjusted value “gradually” diminishes as it deviates farther from the observed value. A large volume of literature is available on interpretations and applications of fuzzy sets and membership functions, including the work of Tanaka [8], Yager and Filev [9], Zimmermann [10], Klir and Wierman [11].

Here triangular-shaped membership functions are assumed, following the discussion made by other authors about the use of full fuzzy linear programming using symmetric triangular fuzzy number [12]. This representation is convenient computationally (a linear program can be used) and is consistent with uncertainty about the “most probable” value. Given an observed value (x_i^{obs}) and its tolerance (α_i) (usually expressed as a percentage of the observed value), Eq. (1) defines the membership function. However, if additional information about the character of the observed value is available, the shape of the membership function could be modified.

$$h_i(x_i) = \max \left\{ 0, 1 - \frac{|x_i - x_i^{\text{obs}}|}{\alpha_i x_i^{\text{obs}}} \right\} \quad (1)$$

x_i is the adjusted value for the i th variable. That is: $\forall i; x_i \in [x_i^{\text{obs}} - \alpha_i x_i^{\text{obs}}, x_i^{\text{obs}} + \alpha_i x_i^{\text{obs}}]$. α_i may have a different value for each observed value, depending on how reliable it is (the less reliable the input data is, the higher it will be).

Cases (a) and (d) in Fig. 1 are specific cases of case (b). A fixed number $\alpha_i=0$ forces its value to be kept after the adjustment, i.e., $x_i^{\text{obs}} = x_i$. In the case of a missing value $h_i(x_i) = 1$ in $(0, L_{\max})$, where $h_i(x_i)$ is the membership grade.

The mathematical problem that needs to be solved in order to find the solution is:

Given a set of observed values $\{x_i^{\text{obs}}\}$ $i \in I_b \cup I_a \cup I_L = I$, (where I is a set of indexes, and I_b , I_a and I_L are the number of boardings, alightings and loads respectively) each with a tolerance of α_i , we define the feasible region as the set $A \subset R^n$, such that $\forall \vec{x} = \{x_i\} \in A$ where the following conditions are satisfied:

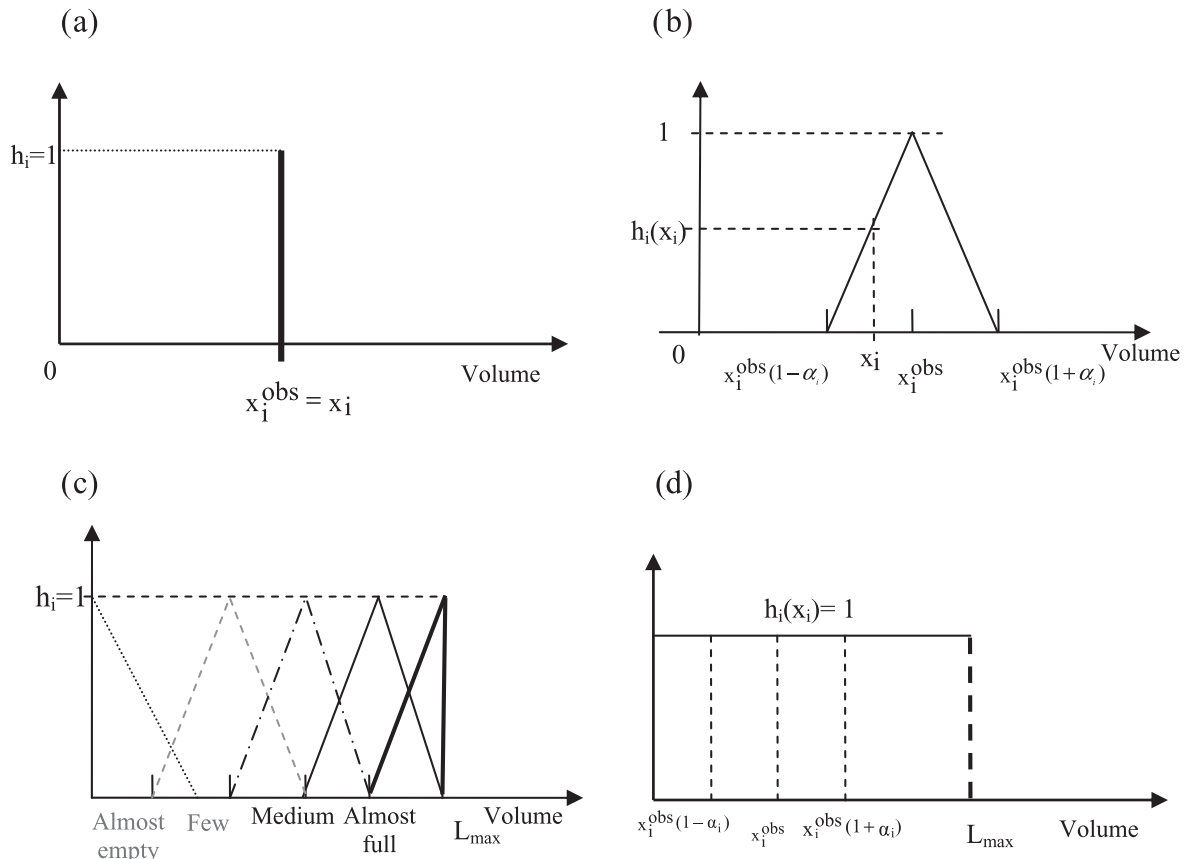


Fig. 1. Membership functions for (a) fix number, (b) crisp number, (c) fuzzy information and (d) missing value.

1. $x_i^{obs} - \alpha_i x_i^{obs} \leq x_i \leq x_i^{obs} + \alpha_i x_i^{obs}$, where x_i^{obs} is the number of passengers who have observed boarding or alighting at stop i and x_i is the adjusted value based on the observed value i .
2. Vector $\vec{x}$ verifies flow conservation law: $\sum_{i \in I_b} x_i = \sum_{i \in I_a} x_i$

Assuming $x_i^{obs} \geq 0, \forall i \in I$, this adjustment becomes a problem of finding out the best solution to the linear optimization problem proposed. The methodology proposed comprises two steps and was already introduced by the authors in De Oña et al. [13]:

Step 1. The problem is solved using MaxMin Method (MM method), (Eq. (2)), and we obtain a value of $h = \min(h_i)$.

$$\text{Max}(h) \text{ where } h \text{ is } \min(h_i) \quad (2)$$

subject to

- Constraints related to the membership functions:

$$h_i(x_i) \geq h \text{ for } i = 1, 2, \dots, 3N \quad (3)$$

where N is the number of transit stops, which means there are $2N + N$ constraints

- Constraints related to the conservation of flow in the transit line:

$$\sum_{i \in I_b} x_i = \sum_{i \in I_a} x_i \quad ; \text{ for } i = 1, 2, \dots, 3N \quad (4)$$

where N is the number of transit stops

- Constraints related to vehicle conditions:

$$l_k \geq 0 \text{ and } l_k \leq L_{\max} \quad (5)$$

where l_k is the number of passengers on board between stops k and $k + 1$ and $L_{\max}$ is maximum vehicle load. Once, the Step 1 is finished, the optimum value for $h = h^*$ is recorded.

Step 2. The problem is solved using the Maximum Sum Method (MS method) (Eq. (6)):

$$\text{Max}(g) \text{ where } g \text{ is } \sum(h_i) \quad (6)$$

Subject to the same constraints related to the conservation of flow at the transit line (Eqs. (4) and (5)), and to the following constraints related to the membership functions:

$$h_i(x_i) \geq h^* \quad \text{for } i = 1, 2, \dots, 3N \quad (7)$$

The total number of unknowns in Step 2 is reduced by one compared to Step 1.

The main difference here with regard to existing models is that now the input data can be qualitative, and the proposed method is able to preprocess them by assigning them a membership function in order to be processed in the same way as the crisp data.

The benefit of transforming the problem into a linear programming problem is being able to count on multiple and optimized routines for the solution [14,12].

4. Data, methodology and statistical analysis

In this section, the proposed method is applied to a real transit line in Malaga to analyze the results. Furthermore, to generalize and validate the results the method is applied to a set of different lines with different number of stops, different boardings, alightings and load data, that have been generated specifically for this purpose. Depending on the amount of qualitative information available, different scenarios are considered and analyzed.

4.1. Example 1: transit line in Malaga

Fig. 2 shows line number 20 in Malaga (Spain). This transit line runs between the City Centre of Malaga (Alameda Principal) to the west area of the city (University). It is 10.6 km long and presents 21 bus stops. Table 1 shows the true boarding and alighting data (True Value, x_i^{true}) for bus number 541. The consistency of the data can be verified: data comply with flow conservation along the line, so the sum of boarding passengers is equal to the sum of alighting passengers on the transit line (Eq. (4)).

From this consistent data we randomly deform values $\pm 25\%$ for the alightings and $\pm 20\%$ for the loads between stops, keeping the boarding fixed. The maximum load for the articulated buses used in this line is 100 passengers.

Having obtained a database within the above-mentioned tolerance, it could then be considered as the data that would be obtained in a counting campaign in which all 60 potential boardings, alightings and loads would be measured. Therefore, it could be taken as the series of observed data in a public transit line (Observed Values, x_i^{obs}). In this case, the values would not be consistent; according to the above definition (the sum of boarding passengers is equal to the sum of alighting passengers on the transit line). In order to state conclusions about the goodness of the method, this process was repeated 1000 times. Therefore, from the true consistent data x_i^{true} (see Table 1), 1000 random databases were generated to be used as the potential observed data in different tours of the line or different hourly base.

The fact that a base of consistent data is used and subsequently randomly distorted allows verifying the goodness of fit of the proposed method.

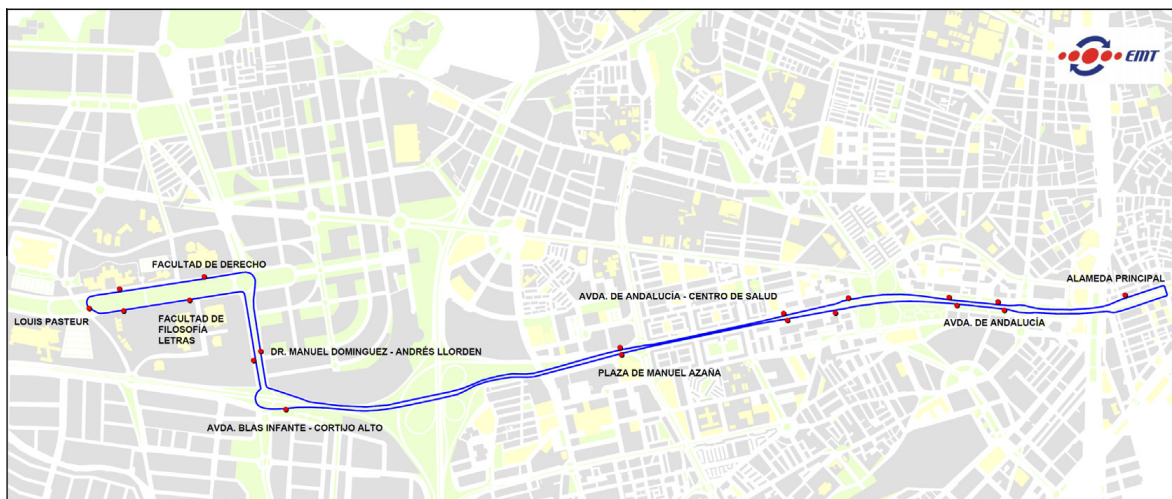


Fig. 2. Example of a transit line in Malaga.

Table 1Alightings and boardings true values for a transit line in Malaga^{*}.

Stop	Stop ID EMTSAM	Boarding	Alighting	Load
1	2301	45	0	45
2	2009	17	2	60
3	1403	4	3	61
4	1404	10	2	69
5	1405	4	1	72
6	2003	15	2	85
7	2007	0	11	74
8	833	0	24	50
9	818	0	37	13
10	2056	0	8	5
11	2056	2	2	5
12	850	0	1	4
13	832	4	3	5
14	2058	1	2	4
15	2059	2	0	6
16	2055	6	1	11
17	1460	1	3	9
18	1461	3	0	12
19	1462	0	2	10
20	1463	0	3	7
21	2301	0	7	0

^{*} EMTSAM is the Public Transport Company in Malaga Municipality.

4.2. Examples for validate and generalize the results

In order to verify that the results obtained can be generalized to any transit line, the method is also applied to a set of 50 different lines, where the number of stops is chosen within the range (10, 75). The procedure was the following:

1. The number of stops is defined and a fictitious transit line is generated with a set of boardings, alightings and loads. Apart from the number of stops, the conditions that boardings, alightings and loads verify the constraints related to the conservation of flow (Eq. (4)) and related to vehicle conditions (Eq. (5)) are imposed. This database is used to verify the goodness of fit of the method (see Sections 3.3 and 3.4).
2. In every fictitious transit line, the consistent data generated is randomly deform in the same way and with the same tolerance as it was for the transit line in Malaga (see Section 3.1): $\pm 25\%$ for the alightings; $\pm 20\%$ for the loads between stops, keeping the boardings fixed. These boardings, alightings and loads, do not satisfy the conditions defined by Eqs. (4) and (5), and are considered as the data that would be obtained during a conventional data collection, and they are the input for the model.
3. In the aim of considering different tours of the same line, different hourly or daily volumes along the line, or even different lines; for every fictitious transit line in Step 1 (50 lines) 100 potential boardings, alighting and loads database are obtained.

So, for generalize and validate the proposed method we will apply it to 5000 different transit lines with a number of stops between 10 and 75.

4.3. Scenarios

As pointed in Section 2.1, it is considered that quantitative information on the passengers boarding at all stops is available and these values are assumed to be exact fixed integers. Furthermore, it is considered that quantitative information on the alightings in some of the stops is also available.

Depending on the remaining amount of qualitative and quantitative information available on alightings (A) and on loads (L) different scenarios are considered:

- (a) No further qualitative information is available on the remaining alightings and loads: missing alighting (MA) and missing loads (ML)
- (b) Qualitative information is available on the alightings (FA) where no quantitative information exist
- (c) Qualitative information is available on the vehicle loads (FL) between successive stops
- (d) Qualitative information is available on alightings (FA) and also on vehicle loads (FL).

In the case of the transit line in Malaga 40 scenarios are considered (see Table 2). To analyse the 5000 transit lines for generalization and validation of the method, 12 scenarios are considered (bold scenarios in Table 2).

Table 2
Scenarios definition.

Cases ML/MA	Cases ML/FA	Cases FL/MA	Cases FL/FA
ML/20MA	ML/20FA	FL/20MA	FL/20FA
ML/25MA	ML/25FA	FL/25MA	FL/25FA
ML/30MA	ML/30FA	FL/30MA	FL/30FA
ML/40MA	ML/40FA	FL/40MA	FL/40FA
ML/45MA	ML/45FA	FL/45MA	FL/45FA
ML/50MA	ML/50FA	FL/50MA	FL/50FA
ML/60MA	ML/60FA	FL/60MA	FL/60FA
ML/75MA	ML/75FA	FL/75MA	FL/75FA
ML/80MA	ML/80FA	FL/80MA	FL/80FA
ML/90MA	ML/90FA	FL/90MA	FL/90FA

Note: ML: missing load; FL: fuzzy load; MA: missing alightings; FA: fuzzy alightings.
xxMA: xx% of missing alightings, (100–xx)% of alightings crisp.
xxFA: xx% of alightings fuzzy, (100–xx)% of alightings crisp.

Table 3
Average errors ($n = 1000$) for the 40 different scenarios in the transit line in Malaga.

LSM Case	ε	Cases ML/MA Case	ε	Cases ML/FA Case	ε	Cases FL/MA Case	ε	Cases FL/FA Case	ε
ML/20MA	2.13	ML/20MA	1.90	ML/20FA	0.91	FL/20MA	0.90	FL/20FA	0.82
ML/25MA	2.70	ML/25MA	2.31	ML/25FA	0.97	FL/25MA	1.00	FL/25FA	0.88
ML/30MA	3.16	ML/30MA	2.62	ML/30FA	1.00	FL/30MA	1.11	FL/30FA	0.90
ML/40MA	4.11	ML/40MA	3.28	ML/40FA	1.10	FL/40MA	1.33	FL/40FA	0.96
ML/45MA	4.50	ML/45MA	3.54	ML/45FA	1.14	FL/45MA	1.41	FL/45FA	0.99
ML/50MA	4.96	ML/50MA	3.84	ML/50FA	1.18	FL/50MA	1.52	FL/50FA	1.01
ML/60MA	5.75	ML/60MA	4.37	ML/60FA	1.28	FL/60MA	1.72	FL/60FA	1.07
ML/75MA	6.87	ML/75MA	5.21	ML/75FA	1.35	FL/75MA	2.05	FL/75FA	1.09
ML/80MA	7.24	ML/80MA	5.49	ML/80FA	1.39	FL/80MA	2.17	FL/80FA	1.12
ML/90MA	7.93	ML/90MA	6.02	ML/90FA	1.47	FL/90MA	2.40	FL/90FA	1.20

In Table 2, ML means that all the loads are missing; FL means that we have qualitative information on all the loads; xxMA represents the case that a percentage xx of the alightings are missing; and xxFA represents the case that we have qualitative information about a percentage xx of the alightings. Boardings were considered as fixed data in all cases.

4.4. Statistical Methods

Conventional statistical parameters are used in order to compare the results of the different scenarios such as: average error, standard deviation, minimum and maximum error, and analysis of the variance (ANOVA).

None of the existing methods in the literature is able to process qualitative data for alightings and loads (FA or FL). Therefore, in the scenarios (b), (c) or (d) they miss a lot of information and they are expected to provide worse results. For comparison purposes, we use the Least Square Method (LSM) as benchmark. LSM uses only quantitative data, so it is applied and only compared with the 10 Cases ML/MA (see Table 2).

In both cases under study (one in the case of the transit line in Málaga; and 50 for validation and generalization of the method) the true boarding and alighting data (x_i^{true}) are used as reference to calculate the error that occurs in every database of non-consistent boardings, alightings and loads, in every scenario. Eq. (8) define the absolute error (ε) for the consistent adjusted values (x_i) in relation to x_i^{true} for each line with a certain combination of non-consistent boardings, alightings and loads. ε is defined as the average distance between x_i and x_i^{true} , where n is the number of values observed. ε is calculated using only the alightings, since the boardings were considered to be fixed (i.e., with no errors). If it is capable of obtaining good adjusted values for the alightings, the loads can be obtained by the difference and it can be asserted that the adjustment was good.

$$\varepsilon = \frac{\sum_{i=0}^n |x_i - x_i^{\text{true}}|}{n} \quad (8)$$

The average error, the standard error deviation, the minimum and maximum error can be obtained from ε . Table 3 shows the average errors obtained from ε committed in the 1000 defined cases in Example 1, under the 40 different scenarios. Furthermore, this Table 3 also shows the average errors when LSM is used under the 10 aforementioned scenarios.

Table 4

Results (average error, standard deviation, min, max, and one-factor ANOVA) for three scenarios with LSM and 12 scenarios with the proposed method ($n = 5000$).

	% Crisp alightings	% Fuzzy alightings	Fuzzy loads	No. cases	Average error	Standard dev	Min	Max
Total				75,000	4.86	2.32	1.00	15.33
<i>Least squared method (LSM)</i>								
ML/90MA	10	0	N	5000	9.95 a	1.47	5.76	15.33
ML/60MA	40	0	N	5000	7.98 b	1.35	4.16	14.33
ML/30MA	70	0	N	5000	5.61 c	1.10	2.22	11.89
<i>Proposed method</i>								
ML/90MA	10	0	N	5000	7.59 d	1.01	4.13	12.70
ML/60MA	40	0	N	5000	5.90 e	0.84	2.80	10.78
ML/30MA	70	0	N	5000	4.27 f	0.68	1.67	8.00
ML/90FA	10	90	N	5000	3.45 g	0.53	1.83	6.60
ML/60FA	40	60	N	5000	3.12 h	0.53	1.42	6.10
ML/30FA	70	30	N	5000	2.75 i	0.53	1.00	6.10
FL/90MA	10	0	Y	5000	5.70j	1.11	2.24	10.78
FL/60MA	40	0	Y	5000	4.43 k	0.87	1.52	8.24
FL/30MA	70	0	Y	5000	3.37 l	0.65	1.25	6.24
FL/90FA	10	90	Y	5000	3.06 m	0.49	1.55	5.46
FL/60FA	40	60	Y	5000	2.90 n	0.49	1.45	5.10
FL/30FA	70	30	Y	5000	2.74 i	0.49	1.00	5.10

a, b, c, d, e, f, g, h, i, j, k, l, m, n Denotes differences statistically significant ($p < 0.05$). Two levels with the same letter.

Table 4 shows the average error, the standard error deviation, the minimum and maximum error in the case of validation and generalization of the method. These values are obtained from ε using the 5000 cases under study for the 12 different scenarios. Table 4 also shows the results when LSM is used.

The statistical analysis has been completed by means of analysis of variance (ANOVA), on a quantitative dependent variable (average error) and the independent variables (factors). ANOVA is used to test the hypothesis that several means are not the same. In our analysis we performed one- and two-way ANOVA. In addition to determining that differences between the means exist, several post-hoc LSD tests were considered on factor levels. The factors considered are: for one-way ANOVA, the scenario; and for two-way ANOVA, the percentage of crispy alightings (10%, 40% and 70%), the fuzzy alightings (yes or no) and the fuzzy loads (yes or no). Interactions between factors were also considered, in order to determine if the presence/absence of a factor level increases/decreases the effect on the response variable (average error). Study of Residuals and Bartlett tests were performed for checking assumptions of normality and homoscedasticity, respectively. Calculations were performed using R-statistical program.

5. Results and discussion

The procedure starts using fuzzy functions to code the qualitative information obtained by the analyst or driver. To that end, a fuzzy class and a triangular type membership function is assigned to each one of the qualitative concepts for loads and alightings, and the analyst is asked to provide information according to that coding. Fig. 3 shows the membership functions of the load and of the alightings in a bus carrying 100 passengers.

Table 3 shows the results for the 40 scenarios in the transit line in Malaga. The values in Table 3 represent the average error ($n = 1000$) for each scenario. Fig. 4 shows the results in Table 3 graphically.

Table 3 and Fig. 4 show that:

- The error is gradually lowered in all cases as the percentage of quantitative information on the alightings increases (e.g., for the LSM, the error diminishes 50% when it goes from 10% to 60% of quantitative data on alighting).
- The LSM shows the largest errors for the same level of quantitative information on alightings.
- The more qualitative information is used, the more the average error diminishes (from the ML/MA cases to the FL/FA cases).
- The less quantitative information there is, the greater the effect of qualitative information on the average error. The separation between the curves in Fig. 4 is much greater when only 10% of crisp alightings are available than when 80% are available.

Results in Table 3 show that the larger errors occur when LSM is used (column 1), followed by the results obtained when the method proposed in this paper is used with no qualitative information available, cases ML/MA (column 2). The smallest errors are committed when the proposed method is used with qualitative information available on alightings and on loads, cases FL/FA (column 5). However, the results for the remaining cases, (where only qualitative information is available on the

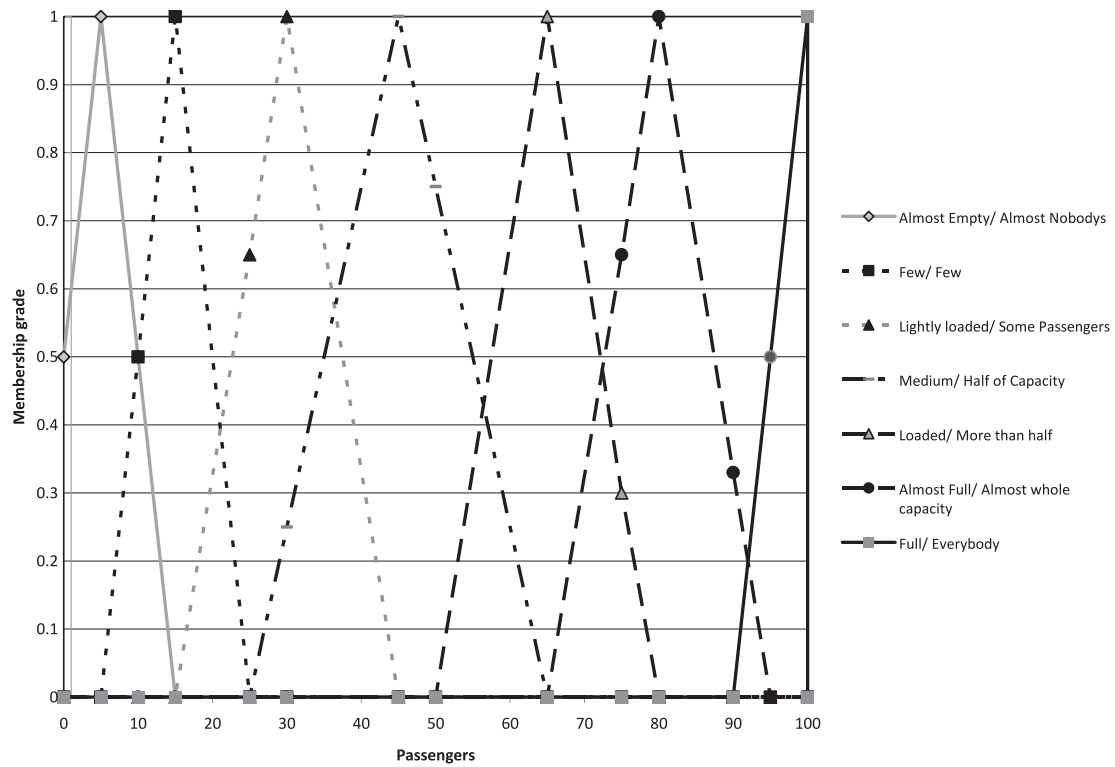


Fig. 3. Membership functions of loads and alightings in a transit line.

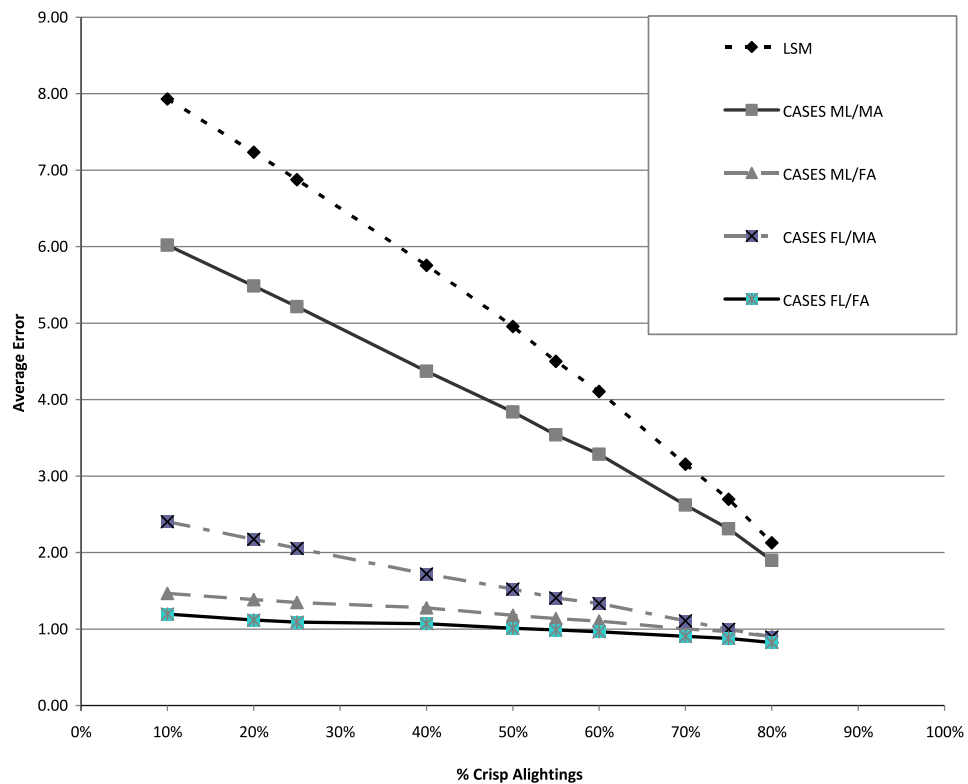


Fig. 4. Average error evolution in the case of a transit line in Malaga.

alightings, cases ML/FA, or on the loads, cases FL/MA) are very similar and results are not conclusive based on the analysis of just one transit line.

Table 5
Results of two-factor ANOVA for the proposed method.

	No. cases	Average error
Total	60,000	4.11
<i>Fuzzy alightings</i>		
NO	30,000	5.21 a
YES	30,000	3.00 b
<i>Fuzzy loads</i>		
NO	30,000	4.51 a
YES	30,000	3.70 b
<i>Crispy alightings</i>		
10%	20,000	4.95 a
40%	20,000	4.09 b
70%	20,000	3.28 c
<i>Fuzzy alightings/fuzzy loads</i>		
NO/YES	15,000	4.50
NO/NO	15,000	5.92
YES/YES	15,000	2.90
YES/NO	15,000	3.10
<i>Fuzzy alightings/crispy alightings</i>		
NO/10%	10,000	6.65
NO/40%	10,000	5.17
NO/70%	10,000	3.82
YES/10%	10,000	3.25
YES/40%	10,000	3.01
YES/70%	10,000	2.75
<i>Fuzzy loads/crispy alightings</i>		
NO/10%	10,000	5.52
NO/40%	10,000	4.51
NO/70%	10,000	3.51
YES/10%	10,000	4.38
YES/40%	10,000	3.67
YES/70%	10,000	3.06

a, b, c Denotes differences statistically significant ($p < 0.05$).

Table 4 shows the results based on the analysis using 50 different transit lines with a number of stops ranging from 10 to 75. 100 possible combinations of non-consistent boardings, alightings and loads have been used for each one of the 50 lines. These combinations data have been adjusted by using the proposed method ($n = 5000$) under the 12 scenarios considered (bold scenarios in Table 2). In order to compare the results, the three scenarios that do not consider qualitative information have been adjusted by using the LSM. 15 cases are compared in total (three scenarios with LSM and 12 scenarios with the proposed method).

In global terms, for all cases ($n = 75,000$), the average error is 4.86, the standard deviation is 2.32, and the minimum and maximum errors are 1.00 and 15.33, respectively.

When the same percentage of crisp alightings is considered (10%, 40% or 70%), LSM produces larger average error, standard deviation, minimum and maximum error. The average error ranges from 5.61 for 70% of crisp alightings to 9.95 for 10% of crisp alightings. The standard deviation ranges from 1.10 to 1.47 (for 70% and 10% of crisp alightings), and the error ranges from 2.22 for 70% of crisp alightings (minimum value) to 15.33 for 10% (maximum value).

From the average error point of view, LSM is followed by the proposed method when no qualitative information is used (ML/MA). The proposed method when only qualitative information on loads is used (FL/MA) is placed the third. In fourth place, when qualitative information on alightings is used (ML/FA) and, finally, the proposed method with qualitative information on both alightings and loads (FL/FA) is the one that produces the smallest average error. For the proposed method, the average error ranges from 2.74 for FL/30FA to 7.59 for ML/90MA; the standard deviation ranges from 0.49 for cases FL/FA to 1.11 for FL/90MA; and the error ranges from 1.00 for ML/30FA and FL/30FA (minimum value) to 12.70 for ML/90MA (maximum value).

The LSD test shows that the scenario has a statistically significant ($p < 0.05$) effect on the average error. 14 different groups were identified (almost one group for each one of the 15 cases being compared). Only the scenarios ML/30FA and FL/30FA show homogeneous groups.

Table 5 shows the two-factor ANOVA results. For this analysis LSM results are not considered. Table 5 shows factors' effect when they are considered in isolation (fuzzy alightings, fuzzy loads and crispy alightings) and the interactions between factors (fuzzy alightings and fuzzy loads, fuzzy alightings and crispy alightings, and fuzzy loads and crispy alightings).

Table 5 shows that when qualitative information is used on the alightings the average error is reduced by an average of 42% by using the proposed method in both cases. When this qualitative information is not used, the average error

($n = 30,000$) is 5.21 whereas if this information is used the average error is 3.00. The LSD test shows that the use of qualitative information on alightings has a statistically significant ($p < 0.05$) effect on average error. The use of qualitative information on the loads between stops reduces the average error an average of 18% (from 4.51 to 3.70). The LSD test also shows that this reduction is statistically significant ($p < 0.05$). Finally, the more qualitative information is available on the alightings, the more the average error diminishes: when qualitative information is increased 30% (from 10% to 40%, or from 40% to 70%) the average error is reduced more than 15%.

When no qualitative information is use on loads and on alightings, the average error is 5.92 when the proposed method is used. This error is lowered in 51% when qualitative information is used on both loads and alightings, reaching an average error of 2.90. When qualitative information is used only on the alightings, the average error is lowered in 48%, reaching an average value of 3.10. These results show that the marginal reduction in the average error when qualitative information on loads is considered is small, (around 24%) with regard to the reduction when qualitative information on alightings is available.

Table 5 also shows that the effect of introducing qualitative information is greater the smaller the quantitative information available. When qualitative information on the alightings is used, the average error is reduced between 28% (from 3.82 to 2.75) and 51% (from 6.65 to 3.25) in the case of 70% of crisp alightings available or 10%, respectively. When qualitative information on loads is used, the average error is reduced between 13% (from 3.51 to 3.06) and 21% (from 5.52 to 4.38) in the case of 70% of crisp alightings or 10%, respectively.

6. Summary and conclusions

The number of passengers boarding and alighting at each transit stop is basic information used in the analysis of urban transit buses operations, to get the loads and being able to calibrate the O/D matrix obtained from surveys. However, observed counts of boardings and alightings often do not match, and on the other hand, alighting data are barely available in the actual urban transit buses systems. The literature gives several different methods that are used to adjust boardings and alightings so the basic principles of flow conservation are met. The methods are characterized by the need for numeric information in order to make the adjustment and the fact that the information must be obtained by automated or manual counts. Therefore, the effort tends to be considerable.

In this paper we propose a method that allows adjustments to boardings and alightings in a transit line based on the qualitative information of the driver, observer or analyst's perception of vehicle loads between stops and on the number of passengers who alights at each stop. This information can be obtained at a low cost by public transport companies since by having a quick look of the vehicle, the driver can choose one of the options defined beforehand (empty, almost empty, ...) by the analyst.

The benefits of the proposed method are:

1. It works on those cases where other methods provide no solution, when there are not available means to obtain a value on the passengers who alight at the stops.
2. It enables data adjustments in the cases where counts can be made, but certain data is missing, thereby preventing the need to make a complete measurement of the public transport line all over again.

To validate the proposed method, it was applied to the adjustment of boardings and alightings on a real transit line in Malaga (Spain) for which consistent real data were known. This enabled the simulation of different scenarios of inconsistent data and the error committed in the adjustment could be verified. Furthermore, to generalize the results, the method is applied to a set of 50 different transit lines, with different number of stops and different in-out data.

The main conclusions that can be drawn are:

- Even without using qualitative information on loads and/or alightings, the errors committed by the proposed method are minor compared to the errors committed by the LSM.
- When qualitative information is used only on the alightings, the average error is reduced in more than a 40% with regard to the case when no qualitative information is used.
- When qualitative information is used only on the loads, the average error is reduced in more than a 15% with regard to the case when no qualitative information is used
- So, using qualitative information on alightings can reduce the average error more than using qualitative information on loads.

Finally, error reductions obtained when qualitative information on loads and alightings is used (51% in average) are lightly larger than those obtained when qualitative information only on alightings is used (48% in average). For that reason, results show that if it was mandatory to choose, it is better to use qualitative information on the alightings than on the loads.

From the operation point of view, this paper also presents a new way to obtain the information about loads between stops, in order to regulate the service, improving and adapting it to the demand in the peak times, making it easier to know when additional vehicles are required.

References

- [1] M.-P. Pelletier, M. Trépanier, C. Morency, Smart card data use in public transit: a literature review, *Transp. Res. Part C* 19 (1) (2011) 557–568.
- [2] P. Blythe, Improving public transport ticketing through smart cards, *Proceedings of the Institute of Civil Engineers, Municipal Engineer* 157 (2004) 47–54.
- [3] S.P. Dempsey, Privacy issues with the use of smart cards, *Legal Research Digest*, 2008, p. 25.
- [4] Z. Qing, W. Yingzhe, L. Jiankou, Public transport IC card data analysis and operation strategy research based on data mining technology, *International Forum on Computer Science-Technology and Applications*, 2009
- [5] S. Kikuchi, D. Miljkovic, H.J. Van Zuylen, Examination of methods that adjust observed traffic volumes on a network, *Transp. Res. Rec.* 1717 (2000) 109–119.
- [6] S. Kikuchi, D. Miljkovic, Method to preprocess observed traffic data for consistency: application of fuzzy optimization concept, *Transp. Res. Rec.* 1679 (1999) 73–80.
- [7] H. Chen, Stochastic optimization in computing multiple headways for a single bus line, in: *Proceedings of the 35th Annual Simulation Symposium IEEE*, 2002.
- [8] H. Tanaka, *Fuzzy Modeling and Its Applications*, Asakura Shoten, Tokyo, Japan, 1990.
- [9] R. Yager, D.P. Filev, *Essentials of Fuzzy Modeling and Control*, John Wiley and Sons, New York, 1994.
- [10] H.J. Zimmermann, *Fuzzy Set Theory and Its Applications*, Kluwer Academic Publishers, Norwell, MA, 1996.
- [11] G.J. Klir, M.J. Wierman, *Uncertainty-Based Information Elements of Generalized Information Theory*, Physica-Verlag, Heidelberg, Germany, 1999
- [12] F.H. Lotfi, T. Allahviranloo, M.A. Jondabeh, L. Alizadeh, Solving a full fuzzy linear programming using lexicography method and fuzzy approximate solution, *Appl. Math. Model.* 33 (2009) 3151–3156.
- [13] J. De Oña, P. Gomez, E. Merida-Casermeiro, Bilevel fuzzy optimization to pre-process traffic data to satisfy the law of flow conservation, *Transp. Res. Part C* 19 (1) (2011) 29–39.
- [14] LINPROG. <http://www.mathworks.com/help/toolbox/optim/ug/linprog.html>, Last accessed: March 22, 2011.