

UNIVERSIDAD DE GRANADA
E.T.S. INGENIEROS DE CAMINOS, CANALES Y PUERTOS
DEPARTAMENTO DE INGENIERÍA CIVIL



PROGRAMA DE DOCTORADO:
*“Seguridad, Calidad y Optimización de Recursos en Infraestructuras y su
Relación Medioambiental (242.99.2)”*

TESIS DOCTORAL

**TRATAMIENTO DE LA INCERTIDUMBRE EN LOS
PROBLEMAS DE PLANIFICACIÓN DEL TRANSPORTE
UTILIZANDO LÓGICA DIFUSA**

**FUZZY LOGIC METHOD TO DEAL WITH
UNCERTAINTY IN TRANSPORT PROBLEMS**

AUTOR:
PENÉLOPE GÓMEZ JIMÉNEZ

DIRECTORES:
JUAN DE OÑA LÓPEZ
Doctor Ingeniero de Caminos, Canales y Puertos

ENRIQUE MÉRIDA CASERMEIRO
Doctor Ingeniería Informática

Editor: Editorial de la Universidad de Granada
Autor: Penélope Gómez Jiménez
D.L.: GR 1830-2014
ISBN: 978-84-9083-013-0

TESIS DOCTORAL-2013

Memoria presentada por Penélope Gómez Jiménez para aspirar al grado de Doctor por la Universidad de Granada con mención de Doctorado Internacional.

La doctoranda, Penélope Gómez Jiménez, y los directores de la tesis “Fuzzy Logic Method to deal with uncertainty in Transport Problems”, garantizamos, al firmar esta tesis doctoral, que el trabajo ha sido realizado por la doctoranda bajo la dirección de los directores de la tesis y hasta donde nuestro conocimiento alcanza, en la realización del trabajo, se han respetado los derechos de otros autores a ser citados, cuando se han utilizado sus resultados o publicaciones.

Granada, Noviembre de 2013

Director/es de la Tesis

Fdo.: Juan de Oña López

Fdo.: Enrique Mérida Casermeiro

Doctoranda

Fdo.: Penélope Gómez Jiménez

|

*"Life is the art of drawing sufficient
conclusions from insufficient premises."
Samuel Butler*

ACKNOWLEDGEMENTS

As I reach the end of this long journey that my doctoral studies have meant for me and all the people close to me, I have to express my gratitude to my advisors, Dr Juan de Oña López and Dr Enrique Mérida Casermeiro, for initiating and watching over my work, for their constant useful remarks and relevant advice, and for always being available. Not only have I been rewarded with their knowledge and feedback, but I have also gained two friends.

I would like to thank Dr Mónica Menéndez for hosting me in the Traffic Engineering Research Group of the Institute for Transport Planning and Systems (IVT), which she heads in the ETH in Zurich, during my stay under the supervision of Sukran Ilgin Guler. This has given me the chance to test the model developed in my thesis in the city of Zurich, and to take it a step further by using it for more applications with very encouraging results, thereby opening up new fields of research.

I have to thank my family for all their support and patience, especially my husband for cheering me up whenever I felt down. Furthermore, I would like to apologise to my children, Rafa and Javi, for having stolen my leisure time with them to focus on my doctorate; I hope that when they grow up they will understand why and realize they gave me the strength to tackle all the challenges of life.

RESUMEN

En esta tesis doctoral, se ha desarrollado un modelo capaz de tratar y manipular información, tanto numérica (aunque pueda provenir de medidas afectadas de errores), como subjetiva, que normalmente viene expresada en términos ambiguos o difusos. Asimismo, el modelo es capaz de imputar valores a los datos perdidos.

Este método, se ha aplicado a diversos problemas relacionados con la Ingeniería Civil y más específicamente, con el área de la ingeniería de los Transportes.

El valor añadido que presenta respecto a los métodos clásicos, donde sólo se permite el uso de información numérica, es que permite añadir información subjetiva procedente del analista. El uso de esta información, de la que usualmente se dispone y que los métodos clásicos desprecian por no ser capaces de tratarla, permite abaratar costes o mejorar la precisión en el ajuste de datos.

Para verificar su utilidad, se ha aplicado a distintos problemas dentro del ámbito de la planificación del transporte. Estos problemas se caracterizan por un gran volumen de datos, alta interdependencia entre ellos, numerosas restricciones, que junto con los métodos de medidas afectados de errores, producen valores observados inconsistentes y, por tanto, tienen que ser preprocesados antes de que puedan ser utilizados en los algoritmos de predicción y toma de decisiones, monitoreo y evaluación en la planificación del transporte para obtener una base de datos consistente.

Concretamente, el modelo se ha aplicado al ajuste de datos de aforo en redes de carreteras, y a la detección de las estaciones aforo que están fallando. Los trabajos desarrollados durante la elaboración de la tesis tratan también de resolver uno de los principales problemas de operación en la planificación de transporte público: el ajuste de pasajeros que suben y bajan en una línea de transporte público.

Las principales aportaciones de esta tesis a través del modelo creado son:

- 1) La obtención de un conjunto valores ajustados y consistentes preservando la integridad de los valores observados, gracias a la posibilidad de incluir la percepción subjetiva del analista.
- 2) La detección de las estaciones de aforo que están fallando sin necesidad de disponer de información adicional.
- 3) La imputación de valores a los datos recogidos en campo, cuya información se ha perdido.
- 4) El ajuste de viajeros que suben y bajan en una línea de tránsito de transporte público, sin necesidad de realizar conteos en todas las paradas.

Los beneficios del método propuesto son diversos. En primer lugar, que funciona en los casos en que otros métodos no proporcionan ninguna solución, porque no disponen de medios para obtener un valor numérico de los datos de campo, ya sea el volumen de tráfico, o los pasajeros que se bajan en las paradas. En segundo lugar, permite obtener los datos ajustados incluso en los casos en los que aún habiéndose realizado los conteos, se ha perdido información, evitando así la necesidad de repetir la colecta de los datos de campo.

Para resumir, en esta tesis se presenta un potente método, que se puede aplicar a diferentes problemas de transporte que tienen que tratar con datos en los que subyace la incertidumbre y la ambigüedad, pudiendo diferenciar entre datos fiables y datos poco fiables.

Esto último es muy alentador ya que, como se verá en las futuras líneas de investigación, se está empezando a aplicar en la predicción de matrices O/D, y tiene un alto potencial para poder ser aplicado en muchos otros problemas donde existe incertidumbre en el transporte.

ABSTRACT

In this PhD thesis a model is developed to process and handle both numerical information, including that derived from error-affected measurements, and subjective information, which may be expressed in ambiguous or vague terms. The model is also able to impute values for missing data.

This method has been applied to various problems related to Civil Engineering and, more specifically, in the area of Transport Engineering.

In contrast to classical methods, which only accept numerical data, the novelty of this model is that it can include subjective information from the analyst. Such information may reduce costs or improve the accuracy of the adjustment data.

To verify its usefulness, the model has been applied to various problems in the field of transport planning. These problems are characterized by a large volume of highly interdependent data and numerous constraints, which together with error-prone measuring methods produce inconsistent observed values. Such data therefore need to be pre-processed in a way that will make them consistent before they can be used in algorithms for prediction, monitoring and decision-making purposes.

Specifically, the model has been applied to produce a consistent set of traffic volume data, which arguably are the most important traffic data in a road network, to detect faulty traffic count stations (TCS) and lastly to deal with operational problems in public transport planning by adjusting numbers of boarding and alighting passengers on a transit line.

The main contributions made by this thesis through the model created are:

- 1) Obtaining a consistent set of values while preserving the integrity of the observed values, which may include their reliability as perceived subjectively by the analyst;

- 2) Identifying faulty TCS without the need for additional information;

3) Imputing values for data collected in the field whose information is lost;

4) Adjusting the numbers of boarding and alighting passengers on a public transport transit line, without requiring counts at all stops.

Among the numerous benefits of the proposed method, two stand out. Firstly, it works in situations where other methods provide no solution, when no means are available to obtain a numeric value for field data, whether traffic volume or numbers of alighting passengers. Secondly, data can be adjusted in those cases where counts can be made but certain data are missing, thereby avoiding the need to measure the field data all over again.

To sum up, a powerful method is presented that can be applied to different transport problems involving data in which there is uncertainty or ambiguity, by taking into account the analyst's subjective perception of the reliability of the data.

This is very encouraging because, as shown in the section on future research, the model has started to be applied in the prediction of origin-destination matrices, and has the potential to deal with the uncertainty that exists in many other transport problems.

Table of contents

ACKNOWLEDGEMENTS.....	ii
RESUMEN	iv
ABSTRACT	vi
LIST OF TABLES	x
LIST OF FIGURES.....	xiv
1 Chapter 1: INTRODUCTION.....	1
1.1. Presentation	1
1.2. Statement of the problem.....	3
1.3. Objectives and Research Hypotheses	5
1.3.1. General Objective	5
1.3.2. Specific Objectives	6
1.3.3. Research hypotheses.....	7
1.3.4. Justification and structure of the thesis.....	8
1.4. Thesis Organization	9
2 Chapter 2: THEORETICAL AND METHODOLOGICAL BASES	11
2.1. Methods to adjust and optimize field data to reach consistency.....	12
2.1.1. Fuzzy-Optimization Method	15
2.1.2. Fuzzy logic applied to transportation engineering.....	19
2.2. Detection of malfunctioning traffic count stations	24
2.2.1. The problem of missing data: Imputation data.....	29
2.2.2. Detecting erroneous data	31
2.3. Adjustment boarding and alighting passengers on a bus transit line using qualitative information	33
3 Chapter 3: MATERIALS AND METHOD	39
3.1. Data.....	40
3.2. Relationships among data.....	42
3.3. Optimization criteria.....	43
3.4. Results	46
3.5. Applications.....	47
4 Chapter 4: ARTICLES	49
4.1. Introduction	49
4.2. Bilevel fuzzy optimization to pre-process traffic data to satisfy the law of flow conservation	51
4.2.1. Introduction.....	52
4.2.2. Description of the problem.....	54

4.2.3.	The bilevel fuzzy optimization method	56
4.2.4.	Example network	63
4.2.5.	Results.....	64
4.2.6.	Summary and conclusions.....	69
4.2.7.	Acknowledgements.....	71
4.3.	Method to Detect Malfunctioning Traffic Count Stations.....	72
4.3.1.	Introduction	72
4.3.2.	Methodology	75
4.3.3.	Application to an urban network.....	82
4.3.4.	Sensibility analysis to different variables	85
4.3.5.	Summary and Conclusions.....	90
4.3.6.	Acknowledgment	91
4.4.	Adjustment boarding and alighting passengers on a bus transit line using qualitative information	92
4.4.1.	Introduction	93
4.4.2.	Background	94
4.4.3.	Theoretical Approach.....	96
4.4.4.	Data, Methodology and Statistical Analysis	101
4.4.5.	Results and Discussion.	107
4.4.6.	Summary and Conclusions.....	113
5	Chapter 5: CONCLUSIONS AND FUTURE RESEARCH	115
5.1.	General Conclusions.....	115
5.2.	Future Research	118
	REFERENCES.....	121
	APPENDIX: PUBLISHED PAPERS	135

LIST OF TABLES

List of tables

Table 1. Detector technologies for traffic counts in a traffic network.....	26
Table 2. Results MM and MS in Simple example of Figure 5	44
Table 3. Comparison of Existing Fuzzy optimization Methods with BO	45
Table 4. Example 1: base data, randomized inconsistent data, adjusted data, and results for different α ranges	55
Table 5. Real intersection in the South of Spain: real base data with missing values, adjusted data, and results for different α ranges	68
Table 6. Results with a simulated error of 75, 50, 25 and 10%.....	84
Table 7. Results for center and edge detectors with a simulated error of 75, 50, 25 and 10%	86
Table 8. Results with a simulated error of 75, 50, 25 and 10% with four, seven and ten not measured movements.....	87
Table 9. Ratios calculated for every scenario providing the standard deviation.....	89
Table 10. Test of hypotheses. Significant cases in bold.	90
Table 11. Alightings and boardings true values for a transit line in Malaga	102
Table 12. Scenarios definition.....	105
Table 13. Results for the 40 different scenarios of 1,000 examples	108
Table 14. Results (average error, standard deviation, min, max, and one-factor ANOVA) for 3 scenarios with LSM and 12 scenarios with the proposed method (n=5,000)	110
Table 15. Results of two-factor ANOVA for the proposed method.....	111

LIST OF FIGURES

List of figures

Figure 1. Typical case of in-flow and out-flow volume inconsistency.	13
Figure 2. Existing methods used to resolve inconsistency of volume flow data.	15
Figure 3. Fuzzy-Optimization Method.	17
Figure 4. Schema of general pre-processing of observed data in a traffic network.....	28
Figure 5. Simple network	39
Figure 6. Membership functions for (a) fix number, (b) crisp number, (c) fuzzy information and (d) missing value	41
Figure 7. Example of situation in which consistent data are available and are randomized to get traffic counts not consistent to explain the theory.....	54
Figure 8. Triangular membership function.....	59
Figure 9. Not consistent real base data set of traffic counts for intersection in Andalusia (South of Spain).....	66
Figure 10. Movements in every node of the network.....	67
Figure 11. Missing values' membership function.....	67
Figure 12. Verisimilitude function for a single observation.....	77
Figure 13. Example of an Urban Network	82
Figure 14. Sensitivity analysis of success versus increase in the number of data not measured.....	88
Figure 15. Membership functions for (a) fix number, (b) crisp number, (c) fuzzy information and (d) missing value	97
Figure 16. Example of a transit line in Malaga	102
Figure 17. Membership functions of loads and alightings in a transit line	107
Figure 18. Error evolution	108

1 Chapter 1: INTRODUCTION

CHAPTER 1

Introduction

1.1. Presentation

Civil engineering is a field that includes numerous other disciplines that produce useful facilities for the human beings, including roads, dams, waste disposal and others that are used in our daily life. Civil engineering is progressing at a fast pace, as are other disciplines.

Civil engineering is considered as the first discipline of the various branches of engineering after military engineering, and includes the designing, planning, construction, and maintenance of the infrastructure. The works include roads, bridges, buildings, dams, canals, water supply and numerous other facilities that affect the life of human beings. Civil engineering is intimately associated with the private and public sectors, including the individual homeowners and international enterprises. It is one of the oldest engineering professions, and ancient engineering achievements due to civil engineering include the pyramids of Egypt and road systems developed by the Romans.

Civil engineering has a significant role in the life of every human being, though one may not truly sense its importance in our daily routine. The function of civil engineering commences with the start of the day when we take a shower, since the water is delivered through a water supply system including a

well-designed network of pipes, water treatment plant and other numerous associated services. The network of roads on which we drive while proceeding to school or work, the huge structural bridges we come across and the tall buildings where we work, all have been designed and constructed by civil engineers. Even the benefits of electricity we use are available to us through the contribution of civil engineers who constructed the towers for the transmission lines. In fact, no sphere of life may be identified that does not include the contribution of civil engineering. Thus, the importance of civil engineering may be determined according to its usefulness in our daily life.

Civil engineering is a multiple science encompassing numerous sub-disciplines that are closely linked with each other. Every sub-disciplines utilizes technical information obtained from numerous other sciences, and with the advancement in all types of technologies, the civil engineering has also benefited tremendously.

Among all these sub-disciplines, this thesis is developed within the field of Transportation Engineering. Transportation Engineering covers aspects of the highway engineering, traffic engineering, transportation and travelling in general.

The importance of transportation engineering has recently escalated, as the daily demands of life in the modern and globalised world is being much and at times overly dependent on an efficient and safe transport system. It is now accepted that an efficient transport system promotes productivity, whilst a poor transport system hampers the economy.

On the social aspect, transport is regarded as an essential ingredient to maintain one's satisfactory life style. This is added by the fact that a gloomy side of transport is also a cause of concerns, which are the environmental effects of travelling and transport in general.

With these concerns, transportation engineering has evolved from being a pure engineering subject to the point where it has to be closer to the social requirement as well as being sensitive to environmental concerns.

In general terms, transportation engineering covers the knowledge of pavement engineering, alignment design, highway design, construction and maintenance for the highway engineering component. The traffic engineering discipline covers issues of traffic characteristics, road and junction capacity, performance levels, traffic management, and congestion management. Other fields include public transportation, safety, environmental issues, travel behavior and the intelligent transport system.

Many people see Transportation Engineering as an essential motivator for upholding the sustainable development concept. A new dawn for transportation is the Intelligent Transport System (ITS), which is defined as a collection of products and systems which utilizes the state of the art technologies in IT, communication, electronics and control to help ensure an efficient, safe and environmentally friendly transportation system

1.2. Statement of the problem

Information on traffic flows between specific origins and destinations in a road network is the main kind of information required by planners and engineers for effective traffic planning, management and control. Origin-destination (O/D) matrices are of vital importance for transport system planning and design, as well as for analysis, modelling and simulation. They contain information about the spatial and temporal distribution of movements between different traffic zones in an area (i.e. each cell represents the number of trips between an origin and a destination). O/D matrices are used to represent the current demand for transport systems; or, in conjunction with anticipated economic and population growth, land-use changes and planning policies, to identify and forecast future demand and other alternative scenarios.

The methods used to estimate O/D matrices are based on the hypothetical availability of accurate traffic volume data and reliable preliminary O/D data. The input data for most traffic networks, however, are either unavailable or contain measurement errors, as in the case of traffic counts and sensor speed measurements. In the past, certain methods were applied to adjust the observed values so they would comply with flow conservation laws at each

network node, aside from other requirements that values need to meet before they can be used as input data in traffic planning algorithms. These are the so-called classical methods.

Missing data processing is another common issue. When input data are available from all sensors, they often contain errors due to sensor operating faults. Most efforts have focused on processing missing values and on detecting and debugging them. Inconsistencies have been avoided by using redundant or related information. Some classical techniques are mean, median, regression and hot-deck imputation. New techniques based on Artificial Intelligence and neural networks, in particular, are being developed. The aim of this thesis is to propose a method that can pre-process field data to make them consistent, while as far as possible preserving their integrity, and that can include their reliability as perceived subjectively by the analyst. The method is based on fuzzy logic and is intended to optimize the solution obtained. The result is a reliable solution that comes close to the observed values, thereby resolving measurement errors in traffic count stations (TCS). The method is also able to detect which TCS is most likely to be faulty. It also allows field data to be processed when there are missing values.

The planning and analysis of transit and public transport operations is problematic and costly. Most current ticketing methods can be used to record where passengers get on board but not where they alight. Current methods are unable to properly reconcile boardings and alightings based on the available data unless they do alighting counts, which is a costly process. As an extension of the research work developed in this thesis, the fuzzy logic method defined in the first part of the thesis is modified and applied to a transit line: counts are made at fewer stops and fuzzy information on alightings and/or vehicle loads between consecutive stops are used to make the boarding and alighting adjustment. Fuzzy information can be obtained from the vehicle driver or an on-board observer, which makes it less costly than the counting method. The proposed method has two main benefits: first, it works in those cases where other methods provide no solution because there is no available means to count the number of passengers who alight at each stop; and, secondly, it makes it possible to adjust data in cases where counts can be made but certain data are

missing, thereby avoiding the need to measure the data on the public transport line all over again.

1.3. Objectives and Research Hypotheses

1.3.1. General Objective

The purpose of this thesis is to address the possibility of using fuzzy set theory in solving complex traffic and transport engineering problems.

Decisions to be made in traffic planning, transport organization and transport management usually need a large volume of input data, which, depending on the context of the problem, include travel time, travel costs, number of vehicles, number of transport facilities, number of passengers, etc. All these data must be sufficiently accurate to help the analyst to understand their dependence on each other. In some situations, very accurate input data are available and, by using a suitable decision-support model, the designer is able to find very satisfactory solutions.

Unfortunately, sufficiently accurate input data are often not available. The input data needed to make decisions are often surrounded by uncertainty.

This uncertainty is especially significant when traffic volume data are involved in the traffic analysis because they generally come from field measurements. Traffic counters and speed sensors are normally used to collect the data, and a variety of factors, including measuring equipment malfunction, unexpected behaviour of a large number of vehicles and human error, will generally make field data inconsistent in many ways.

Field data therefore need to be processed in a way that will make them consistent before they can be used in algorithms for prediction, monitoring and decision-making purposes. The methods used to estimate O/D matrices are based on the hypothetical availability of accurate traffic volume data and reliable preliminary O/D data, so pre-processing the data to ensure it is consistent is essential in order to provide a reliable dataset for making decisions.

The main general objective of this research work is to create a model capable of dealing with the uncertainty underlying any transport problem with a view to producing an O/D matrix that contains accurate and reliable preliminary traffic volume data in the specific case of road networks, or data on alighting and boarding passengers on a public transport route.

1.3.2. Specific Objectives

This thesis focuses on proposing a new method for pre-processing field-collected data to achieve consistency and, whilst pursuing this aim, extending the model to detect malfunctioning TCS.

Arrays of different traffic and transport parameters are all clearly characterized by uncertainty, subjectivity, imprecision and ambiguity. Thus, in the mathematical modelling of traffic and transport processes in which the individual parameters are uncertain, ambiguous or subjectively estimated, mathematical methods must be used that can deal satisfactorily with that uncertainty, ambiguity and subjectivity. Fuzzy set theory is a very convenient mathematical tool for treating these problems.

The existing methods for detecting malfunctioning TCS are reviewed. In these methods the data used for correction always belong to actual historical datasets which are correlated with the erroneous data either temporally or spatially. A problem occurs, however, when no historical data are available, either because they were not measured or because they are missing. To solve this problem, a method for the automatic detection of malfunctioning TCS in a transport system is also presented.

In the planning of public transport networks, it is crucial to know the real O/Ds of passengers, and also the vehicle loads during service operations, in order to decide if an additional vehicle is required because the maximum load has been exceeded, thereby helping to adapt the service as closely to demand as possible.

Obtaining data to use in the planning and analysis of an urban public transport operation is problematic, particularly for urban bus routes. Most urban

bus ticketing methods can be used to record passengers getting on board but not those getting off, and current methods are unable to properly adjust boardings and alightings based on the available data, unless they include alighting counts. To solve this problem, this thesis proposes a method whereby counts are made at fewer stops and qualitative information on alightings and/or vehicle loads between consecutive stops is used to adjust the boarding and alighting figures, prior to obtaining the real O/D of passengers and calibrating the O/D matrix by using the loads between stops. Qualitative information can be obtained from the vehicle driver or an on-board observer, avoiding the need to count many stops during the planning period.

1.3.3. Research hypotheses

The assumptions on which this research is based are presented below and are closely related to the objectives:

Since numerical information is very expensive and sometimes difficult to obtain, is it possible to make use of subjective information to deal with the uncertainties in the field data? Indeed it is, but classical methods cannot be used to process this kind of information, so a more appropriate mathematical tool must be used, such as fuzzy set theory.

After a review of the fuzzy logic methods that have already been applied to transport engineering, a new model based on fuzzy logic is created to adjust the data. However, does this model achieve a better adjustment of field data than the classical ones? In fact, the classical methods are not able to deal with subjective information so they cannot provide a solution. This is shown in this thesis.

Existing methods for detecting malfunctioning TCS use historical data to identify the error, but since such historical data is not always available, the proposed model detects faulty TCS on the assumption that no additional information is available.

In public transport planning, counting alighting passengers is expensive and the collected information is very limited because such fieldwork is only

conducted for a short period of time and only on a sample of the transit lines in the city under study. Could subjective information on the loads and alightings between consecutive stops be used to adjust the boarding and alighting figures? This could be done using fuzzy logic optimization. The model created deals with this adjustment problem and allows the analyst to use subjective information on loads and alightings to forecast the number of passengers alighting at every stop.

1.3.4. Justification and structure of the thesis

The first stage of this research was to create the model to be applied to all these specific transport problems. Once the model had been created, it was applied to three different but related issues. Three papers have been published in three journals indexed in the Journal Citation Reports. The above justifies that this PhD thesis is presented as a group of published articles.

Methods to Adjust and Optimize Field data to reach consistency

Traffic data obtained in the field usually have some errors. For instance, traffic volume data on the various links of a network must be consistent and satisfy flow conservation, but this rarely occurs. This paper presents a method for using fuzzy optimization to adjust observed values so they meet flow conservation equations and any consistency requirements. The novelty lies in the possibility of obtaining the best combination of adjusted values, thereby preserving data integrity as much as possible. The proposed method allows analysts to manage field data reliability by assigning different ranges to each observed value. The paper is divided into two sections: The first section explains the theory through a simple example of a case in which the data is equally reliable and a case in which the observed data comes from more or less reliable sources, and the second one is an actual application of the method in a freeway network in southern Spain where data were available but some data were missing.

Detection of malfunctioning traffic count stations

This study presents a method for the automatic detection of malfunctioning traffic count stations (TCS) in a transport system. First, double linear optimisation is used to detect inadmissible errors in the recordings of a series of TCS and next, the TCS that are most likely to be failing are identified. The method has been applied to an urban traffic network showing success rates up to 93% in identifying the TCS that are failing.

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

Obtaining data to use in an urban public transport operation planning and analysis is problematic, specifically in urban bus transit lines. 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 O/D of passengers allowing the O/D matrix calibration by using the loads among 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 transit line in Malaga (Spain) and to a set of 50 different 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%.

1.4. Thesis Organization

This thesis consists in five chapters:

Chapter 1 includes an introduction to the thesis, a brief description of the general and specific objectives, the research assumptions and the justification of the structure of this document as a group of published papers.

Chapter 2 presents an overview of the theoretical and methodological bases, with reference purpose. This chapter is divided in three subsections, where the theoretical bases of the three legs of this thesis are exposed.

Chapter 3 presents the materials and method, which constitutes the main contribution of this doctorate.

In this chapter the model is explained with an example, to verify its goodness. Once it is demonstrated that the new model leads to better adjustment than the existing methods and allows dealing with uncertainty by using fuzzy information, it will be applied to a road network, to detect malfunctioning TCS and to adjust boardings and alighting passengers in a transit lines. These applications have been published in three different scientific and indexed journals. These articles are shown in **Chapter 4**, which is divided into three subsections and presents the three published papers.

Chapter 5 presents the major conclusions of this work and future research lines.

Finally, we include all the **references** used in this thesis.

The final version of the published papers is in the **Appendix**.

2 Chapter 2: THEORETICAL AND METHODOLOGICAL BASES

CHAPTER 2

Theoretical Preliminaries

The first step in any thesis, after selecting the central theme of the study, is to conduct a thorough review of the existing bibliography in order to establish the theoretical and methodological bases to support the subsequent phases of the research. The approach and development of the methodology proposed here were reached after reviewing the main studies on pre-processing field data to satisfy the law of flow conservation.

This chapter is organised in three sections. First section, contains a review of existing method to pre-process traffic data, and in the second section, there is a brief review of the existing literature for the detection of malfunctioning traffic count stations (TCS) in a transport system.

These two researching issues far from be independent, are closely related. It can be stated that generally, pre-processing of a traffic data set obtained in the field must include two phases:

On one side, traffic volume data on the various links of a network must be consistent and satisfy flow conservation, but this rarely occurs. This chapter presents an overview and a discussion of the existent methods proposed to

optimize the pre-processing of this data, in order to obtain a reliable set of traffic data.

But also, it is important to be able to detect a TCS that is not working properly so its measured values must be rejected or pre-process separately. This is the subject of the second part of the thesis, so this chapter includes a description of the existing methods developed for the detection of malfunctioning traffic count stations.

Finally, there is a third section related to passengers' adjustment in an urban public transport system. In particular, the subject is about the problem of obtaining data to use in an urban public transport operation planning and analysis, particularly in urban bus transit lines. Nevertheless, although the aim is different in this third case, the model used is the same by spreading the aforementioned to be able to adjust the in and out passengers in a transit line.

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.

Therefore, this chapter contains a third section to show a literature review about obtaining and pre-processing data to use in an urban public transport operation planning and analysis.

2.1. Methods to adjust and optimize field data to reach consistency

The methods used to estimate Origin–Destination (O/D) matrices are based on the hypothetical availability of accurate traffic volume data and reliable preliminary O/D data. However, it is very common that the input data for most traffic networks are either unavailable or contain measurement errors, particularly in the case of traffic counts and sensor speed measurements. In fact, some studies (e.g., Zhong et al., 2004) demonstrate that 50% of the permanent traffic counts set up on highways contain missing data, making it difficult to ignore measurement errors when processing data used to plan, design, control and manage traffic (Sharma et al., 1996). So the analysis of a

traffic network becomes a big deal because of the existence of errors and data with a high level of uncertainty. That is why many authors have developed methods for pre-processing data of traffic networks in order to obtain an accuracy level of the data, which allows a proper decision-making.

To illustrate this problem, Figure 1 shows a typical case of non consistency in field data, due to multiple factors that will be discussed all along this paragraph. Arrows in Figure 1 mean observed traffic volumes that must be adjusted to fulfil the requirements of flow conservation. That is to say, in a general traffic network, the sum of in-flows volumes in a node must be equal to total out-flows volumes of the same node.

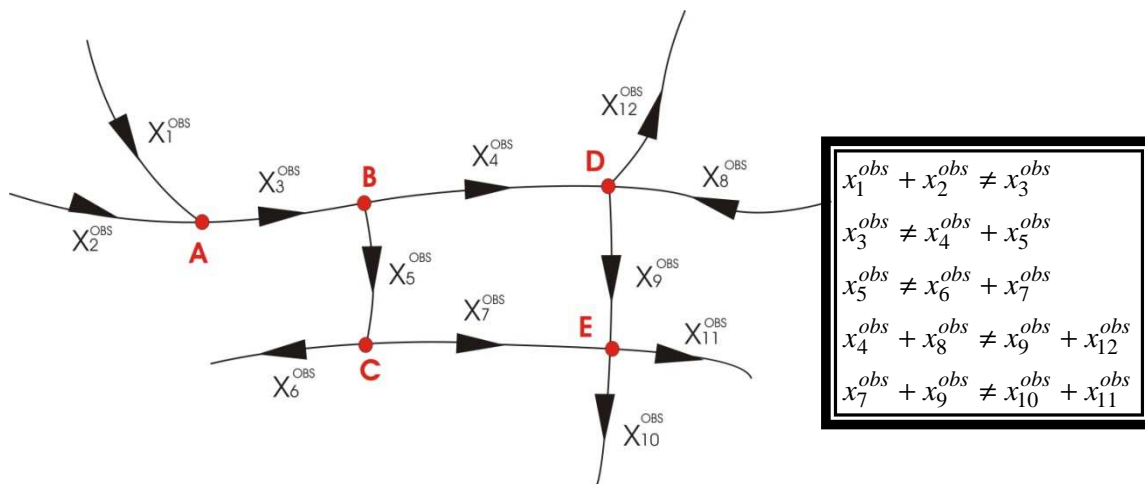


Figure 1. Typical case of in-flow and out-flow volume inconsistency.

Several well-known methods have been developed in order to solve this inconsistency of data. Ideally, any method should meet the following requirements (Kikuchi et al., 2000):

- Ensures the flow consistency at any point (or node) of the network;
- Preserves the integrity of the observed values as much as possible and also incorporates the analyst's knowledge about the differences in accuracy and reliability among the observed values;
- Is able of incorporating the analyst's knowledge about the relationships among the volumes;

- Possesses the logical base for the calculations and a measure that indicates the reasonableness of the adjusted values;
- Handles a large, complicated network with simple computation;
- Is consistent with the structural properties of the data.

Existing methods used for pre-process field volume data are divided into two groups, except for manual method¹, attending on the treatment and characterization of the obtained values.

Group 1 are classic methods and consider that every observed measure value is exact but they may be erroneous. So the analytical framework used in this group is therefore the same as the one used in the statistical-regression analysis. In this case, all the constraints that represent the relationship among the volumes are expressed in rigid terms.

Group 2 assumes that an observed value is not exact but it is very close to the real value. So, the adjusted value is located within a range around the observed values. The analyst must determine the size of the range according to the acceptable deviation and the accuracy of counts. Later, each adjusted value is determined inside that range, as close to the observed value as possible, subject to the constraints of flow conservation. In this case, the constraints can be soft, while approximate information about volume relationships is added to the constraints.

Different methods and its classification are summarized in Figure 2.

¹ The manual method tries to adjust values adding or subtracting one number at a time to every measured flow value at nodes. The analyst starts in one arbitrary selected node and continues with adjacent nodes in a way that, at the end of the process, all nodes are balanced in the whole network. This method can be useful for small networks where the analyst has sufficient knowledge about local traffic conditions to validate the reasonableness of the solution. That is because the adjusted values are heavily dependent on the selection of the node from which the computation begins and the process and obtained results can be scarcely controllable.

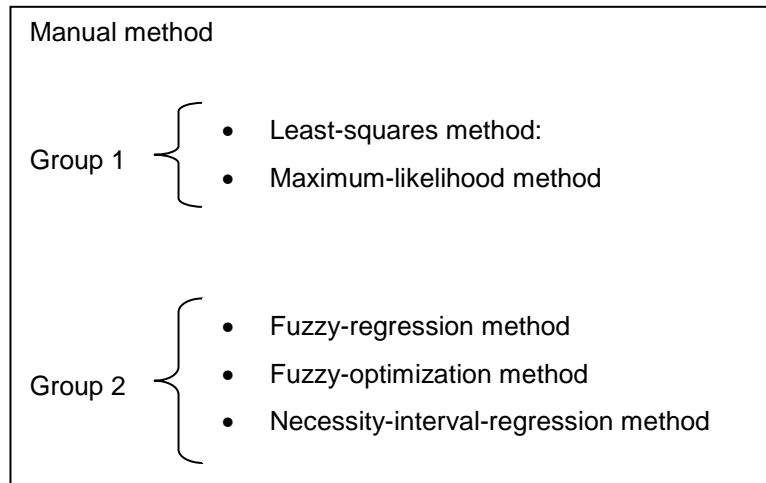


Figure 2. Existing methods used to resolve inconsistency of volume flow data.

A detailed revision of existing methods can be seen in Kikuchi et al., (2000).

Following we will described the Fuzzy Optimization method as it constitutes the theoretical base of our research work.

2.1.1. Fuzzy-Optimization Method

In 1965, Lotfi A. Zadeh published "Fuzzy Sets," which lays out the mathematics of fuzzy set theory and, by extension, fuzzy logic. Zadeh observes that conventional computer logic could not manipulate data that representes subjective or vague ideas, so he creates fuzzy logic to allow computers to determine the distinctions among data with shades of gray, similar to the process of human reasoning.

Although, the technology was introduced in the United States (US), US and European scientist and researchers largely ignored it for years, perhaps because of its unconventional name. They refused to take seriously something that sounded so childlike. Some mathematicians argued that fuzzy logic was merely probability in disguise. But fuzzy logic was readily accepted in Japan, China and other Asian countries. The greatest number of fuzzy researchers today is found in China, with over 10,000 scientists. Japan, though considered at the leading edge of fuzzy studies, has fewer people engaged in fuzzy

research. A decade ago, the Chinese University of Hong Kong surveyed consumer products using fuzzy logic, producing a 100-plus-page report listing washing machines, camcorders, microwave ovens and dozens of other kinds of electrical and electronic products. From the early 70's, the fuzzy logic theory started to be developed, trying to alleviate difficulties in developing and analyzing complex systems encountered by conventional mathematical tools, by observing that human reasoning can utilize concepts and knowledge that do not have well-defined, sharp boundaries.

When solving real-life traffic and transportation problems we should not use only objective knowledge (formulae and equations) or only subjective knowledge (linguistic information). We simply cannot and should not ignore the existence of linguistic information (i.e., subjective knowledge). Fuzzy logic is an extremely suitable tool for combining subjective and objective knowledge.

Traffic planning, transport organization, and traffic and transportation management are processes that are linked to certain decisions that must be made. However, uncertainty often surrounds the input data needed to take those decisions. Thus, in the mathematical modelling phase of traffic and transportation process, whose individual parameters are uncertain, ambiguous or subjectively estimated, mathematical methods used should be able to satisfactorily deal with uncertainty, ambiguity and subjectivity.

The main objective of this thesis is to propose a method for pre-processing field data in a transportation problem, which carries out data's adjustment taking into account subjective information. For that aim, fuzzy logic is applied whose theoretical bases will be exposed all over this chapter.

This procedure considers the notion that each observed data x_i^{obs} is an *approximately* x_i^{obs} as a fuzzy set, that is to say, a fuzzy range around x_i^{obs} , and the adjusted value is expected to be into a range around the observed value. This idea is proposed originally by Kikuchi (1997), and later is extended and improved by Kikuchi and Miljkovic (1999).

In this method there is a membership function of the fuzzy set $[h_{x_i^{obs}}(x_i)]$ that is placed around the observed value, so the adjusted value is located in some place within its base. The analyst must define a membership function for each observed value that represents how close he expects the adjusted value to be found. In other words, the observed value is considered to have the maximum value of membership function and the base width is defined as the acceptable deviation that can have the adjusted value (see Figure 3).

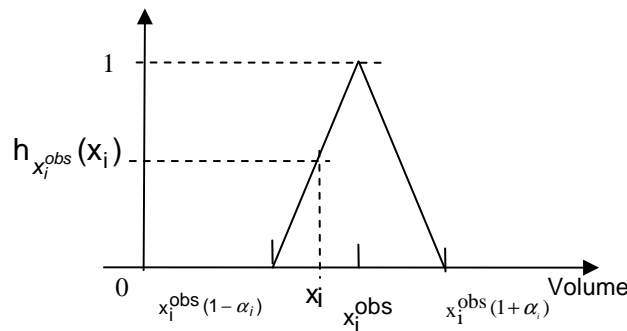


Figure 3. Fuzzy-Optimization Method.

Accordingly to that, for a particular membership function of a fuzzy set around x_i^{obs} , $h_{x_i^{obs}}(x_i)$, the adjusted values $\vec{x} = (x_i)$ can be found applying two methods: by maximizing the minimum $h_{x_i^{obs}}(x_i)$ for all components i 's, and by maximizing the sum of $h_{x_i^{obs}}(x_i)$. Mathematically expressed:

$$\max_{\vec{x} \in A} \min_i \{h_{x_i^{obs}}(x_i)\} \quad (1)$$

or

$$\max_{\vec{x} \in A} \left\{ \sum_i h_{x_i^{obs}}(x_i) \right\} \quad (2)$$

Where A is the feasible region, that is, given a set of observed values $\{x_i^{obs}\}$, $i \in I$, (where I is a set of indexes) each with a tolerance of α_i , we define the *feasible region* as the set $A \subset \mathfrak{R}^n$, such that $\forall \vec{x} = \{x_i\} \in A$ where the following conditions are satisfied:

1. $x_i \geq 0$
2. $x_i^{obs} - \alpha_i x_i^{obs} \leq x_i \leq x_i^{obs} + \alpha_i x_i^{obs}$
3. Vector \vec{x} verifies flow conservation laws

For both cases, the problem is formulated as a linear programming problem.

Fuzzy-Optimization method based on the assumption of maximizing the minimum value of h_i is developed by Kikuchi and Miljkovic (1999). They follow the process of fuzzy optimization principle, first time presented by Bellman-Zadeh (1970) and solve the problem using the fuzzy linear programming method suggested by Zimmermann (1996).

For Fuzzy-Optimization, Kikuchi and Miljkovic (1999) “fuzzify” each observed value considering a fuzzy set with a triangular and symmetric membership function. The membership function needs not to be triangular but if there isn't any other information this is a very reasonable assumption in order to use linear programming formulation because of its straight-line expressions. A nonlinear shape can be used for the membership function, but then, a nonlinear optimization program is needed to find the solutions that will lead to an increase of the computational effort, becoming a problem for larger networks.

Delgado et al. (1992) stated that it makes no sense to use sophisticated shapes for membership functions, taking into account that the linguistic assessments are just approximate assessments, given by the experts and accepted by the decision makers because obtaining more accurate values is impossible or unnecessary. In fact, we consider triangular or trapezoidal membership functions good enough to capture vagueness of linguistics assessments.

In this method, another important step is the selection of the size of the range within the adjusted value that must be found. This range represents the adjusted value's acceptable deviation from the observed data and the analyst has to define it. Anyway, Kikuchi and Miljkovic (1999) prove that the size of the

range has not significant effect on the final adjusted values, but it has to be large enough to find a feasible set of solutions.

In particular, this thesis presents an evolution of this Fuzzy-Optimization Method, so a literature review of this method is shown below.

2.1.2. Fuzzy logic applied to transportation engineering.

The problems in transportation planning and traffic control are frequently ambiguous, vague and deficient in definition and data values reliability; therefore, decisions to be made after a deep analysis of the subject are often characterized by subjectivity.

A variety of complex traffic and transportation problems have been solved using deterministic and stochastic models developed by mathematics based on binary logic. But, since the fuzzy set theory recognizes the vague boundary that exists in some sets of data, fuzzy set theory techniques are highly suitable to be applied to transportation analysis.

As a consequence, there are broad applications on transportation engineering for fuzzy set theory. Since Pappis and Mamdani (1977) first apply fuzzy logic to a transport subject, specifically to traffic signal controllers, many other authors have applied this theory in a wide number of fields within transportation engineering.

In particular, some transportation subjects where fuzzy logic is widely applied are shown below:

Trip generation: Kalić and Teodorović (1997b) solve this problem for the first time applying fuzzy logic techniques, using Wang and Mendel (1992a) procedure. The number of trips for the subsets generated in a given area is estimated by different methods: fuzzy logic, artificial neural networks and multiple linear regressions, resulting that the fuzzy logic approach gives the closest estimate. Also, is in this context where the origin-destination estimation from link counts is settled, being Xu and Chan (1993a, 1993b) who first use fuzzy set theory to analyse the problems arising from the poor quality of link count data. Later, Kikuchi and Miljkovic (1999) develop an improved technique

using fuzzy linear programming method suggested by Zimmermann (1996) that serves as a base for this thesis.

Trip distribution: Kalić and Teodorović (1996, 1997a) use fuzzy logic to estimate the number of air passengers travelling between major industrial cities and given regions and, compared to other non-fuzzy methods, it gives the best approach to the problem.

Modal split: Based on the results obtained by Teodorović and Kalić (1996), Quadrado and Quadrado (1996) use fuzzy logic to establish the accessibility of several transportation modes in the Lisbon Metropolitan Area. They consider that all variables used in the “classical” method for calculating accessibility are characterized by fuzziness, so they develop a fuzzy rule base for each transportation mode.

Route choice: Teodorović and Kikuchi (1991) study the binary route choice problem using fuzzy inference techniques. Akiyama et al. (1993) also develop a model for route choice behaviour based on the fuzzy reasoning approach. Also, Lotan and Koutsopoulos (1993a, 1993b) study models for route choice behaviour in the presence of information based on ideas from approximate reasoning and fuzzy control. This last model has been very important for later researches in Intelligent Vehicle Highway Systems (IVHS). Akiyama and Tsuboi (1996) study route choice behaviour using multi-stage fuzzy reasoning to describe the driver decision-making process on road networks. For that task they consider the multi-route choice problem. They also include a second stage of estimation with a neural network model to represent the number of alternative routes and the values of the utilities of individual alternative routes. After their analysis, they obtain better results with the combination of fuzzy logic (first stage) and neural network (second stage) than using only fuzzy logic for every stage.

Traffic assignment: Akiyama et al. (1994) present a study on the relationship between traffic information and drivers' behaviour. They start from the premise that drivers' perception of time is a triangular fuzzy number and

develop Fuzzified Frank-Wolfe algorithm to design the traffic assignment model on the network of Hanshin Expressway and urban streets in the Osaka area.

Transportation investment project selection: Tzeng and Teng (1993) show the possibilities of using the fuzzy set theory in this field considering a fuzzy multi-objectives problem. Smith (1993) applies it for the evaluation of potential suburban railway station locations on one of three possible railway line extensions to Brisbane's suburban network and including several criteria as population trends or possibility of bus/rail interchange at each possible station location.

Traffic control at intersections: This is the first field in which fuzzy logic was applied. Pappis and Mamdani (1977) develop an approximate reasoning algorithm to control traffic at intersections. For that goal, they assume that vehicles arrive at the intersection with a uniform distribution and suppose that the vehicle detectors are placed upstream from the intersection in a way that is possible to inform the controller about vehicle arrivals at the intersection within the next 11 s. Later, Chang and Shyu (1993) generate a fuzzy expert system to determine whether a traffic signal is needed in an intersection.

Traffic control in a corridor: Nakatsuyama et al. (1983) make a study of a fuzzy logic controller in comparison with a standard vehicle-actuated controller for different values of traffic flow rates. As result, they obtain considerably shorter average delay times using fuzzy logic than using a standard vehicle-actuated controller. Sasaki and Akiyama (1986, 1987, 1988) show that control of an urban expressway depends upon a skilled operator's judgment and decisions, so they describe this operator's judgment process using fuzzy logic. They design a simple fuzzy reasoning model for on-ramp control that is introduced in a model tested on the Osaka-Sakai route of the Hanshin expressway with very reasonably well results. Chen et al. (1990) also develop a model of a fuzzy controller for freeway ramp metering that is tested on the San Francisco-Oakland Bay Bridge. They find that the fuzzy controller is very efficient in reducing efficiency losses due to incidents

Network control: Chiu (1992) uses fuzzy logic to develop an adaptive traffic signal control for small networks of intersections. For this aim, he adjusts independently signal-timing parameters (cycle time, phase split and offset) as functions of the local traffic condition and of the signal timing parameters at adjacent intersections. He develops a fuzzy rule base for the adjustment of every parameter.

Accident analysis and prevention: Akiyama and Shao (1993) investigate the problem of the construction of traffic safety facilities on urban expressways by evaluating costs and benefits from reducing the number of accidents. During the decision analysis, the cost and effectiveness of the traffic safety facilities to be installed are evaluated in monetary terms. But the big problem of the study is that safety costs cannot be defined deterministically, and the cost and benefit of alternatives cannot be measured without fuzziness. In other words, factors such as feeling of safety, driving comfort, etc., must be taken into account when evaluating certain alternatives. The authors use incremental cost-benefit analysis with fuzzy constraints and dynamic programming to solve this problem. This model is tested on the Hanshin Expressway in Japan. Another problem very suitable for fuzzy set theory techniques is the incident detection and the identification of accident-prone locations, studied by Sayed et al. (1995). In addition, Schretter and Hollatz (1996) use fuzzy logic to determine the required period of waiting after a traffic accident when nobody is present at the place of an accident.

Level of service (LOS): Chakroborthy and Kikuchi (1990) apply fuzzy set theory to the analysis of highway capacity and LOS. For the development of the model, they represent the values of input variables (i.e., ideal capacity, sight distance, volume of traffic, and headway between cars) and output variables (i.e., adjustment factors, actual capacity, and LOS criteria) by fuzzy numbers. The authors prove that, if the LOS categories are defined as fuzzy sets, the results are more accurate. Also in this field, Ndoh and Ashford (1994) present a model to evaluate airport passenger services using fuzzy set theory techniques, and Pattnaik and Ramesh Kumar (1996) develop a methodology to define LOS of urban roads taking into account users' perceptions.

Vehicle and crew routing, scheduling and dispatching problems:

Although combinatorial optimization methods or different heuristic algorithms with deterministic characteristics usually solve complex vehicle routing problems, during last years, several authors have published many works for solving these problems with fuzzy set theory. Perincherry and Kikuchi (1990) study the transshipment problem as a fuzzy approach. The transshipment problem has to choose the optimal allocation of goods and services between supply and demand locations taking into account intermediate points where goods can be stored to satisfy the demand at a later date. They plan the problem assuming fuzziness in the quantity of supplies available and the quantities demanded. And consider costs and travel time between the locations as precise information. Another important field for fuzzy theory is the daily planning of transportation companies (Milosavljević et al., 1996). They receive a great number of requests every day from clients wanting to send goods to different destinations. Every request is characterized by the type of freight, the amount of freight (weight and volume), the loading and unloading sites, the preferred time of loading and/or unloading and the transportation distance. So, they have to choose type of vehicles, routes, etc., taking into account the total number of available vehicles, the number of vehicles temporarily out of order, and vehicles undergoing technical examination or preventive maintenance work. The authors prove that using fuzzy logic for this decision problem provides better results in terms of number of ton-kilometres transported.

Air transportation: Larkin (1985) make the first application of fuzzy logic in the field of air traffic control by developing a model for an autopilot controller based on fuzzy logic. Teodorović and Babić (1993), by contrast, apply fuzzy logic for air traffic flow management, including factors concerning congestions at airports. On the other hand, Teodorović et al. (1994) solve the airline network design problem using fuzzy logic and fuzzy mathematical programming.

River transportation: In relation to the process of transporting bulk freight in river traffic, Vukadinović and Teodorović (1994) develop an approximate reasoning model to control the process of loading, transporting and unloading gravel since a dispatcher managed this procedure under high uncertainty conditions.

According to this discussion about fuzzy logic applications in transportation engineering, it could be concluded that almost every problem associated with a degree of uncertainty and/or lack of accuracy may be formulated with a fuzzy logic approach obtaining, in most of the cases, good results.

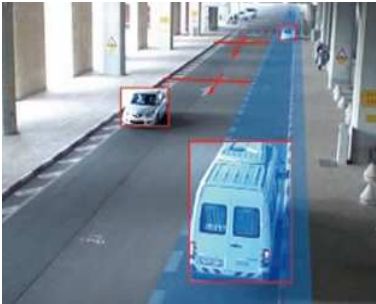
2.2. Detection of malfunctioning traffic count stations


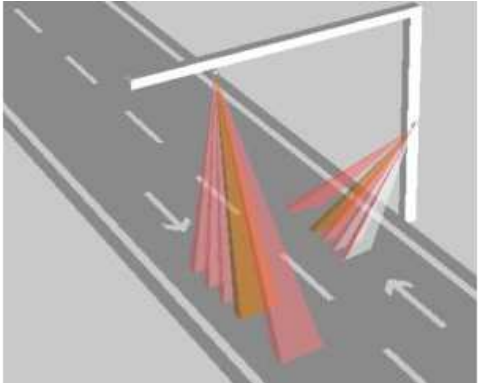
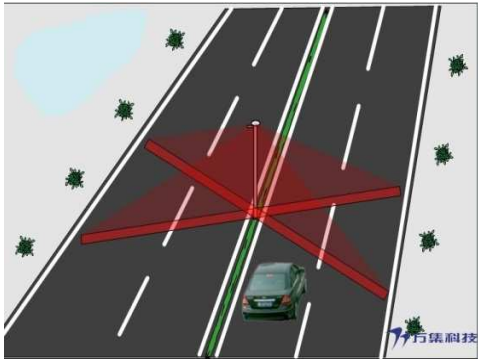
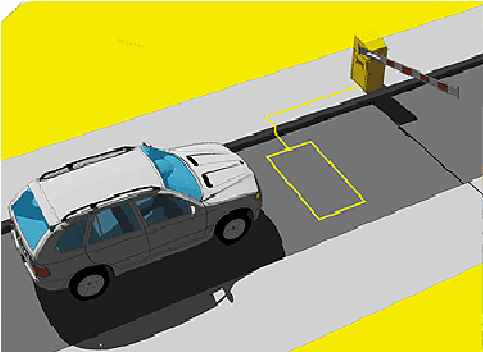
As mentioned before, transport network models need great amount of data to analyse traffic flows in order to make decisions to improve transport in a certain area.

The accuracy of available data will have a decisive influence in the analysis' results and therefore, in the decisions. So, a correct pre-processing of data is very important at the beginning of the transport model construction.

Nowadays, flow data comes from measures carried out on the field using TCS's located at different places all along the transport network. TCS's are integrated in an Intelligent Transportation System (ITS) that is operated by a Traffic Management Centre that maintains control and communication links and also assumes responsibility for archiving the data and performing any quality control measures specified by agency policy.

These systems are very usual in big cities with complex transport infrastructures, but they are not so common in rural areas. There are many detector technologies for field data collection. They are very different and are used depending on the magnitudes to be measured. Table 1 shows the detectors most often used in traffic counts stations (TCS).

Technology	Image	Brief description
Video Detection Technology		Cameras record images of traffic conditions, which can be used to extract several kind of information

Technology	Image	Brief description
<p>Radar and Acoustic Traffic Sensor</p>		<p>The device emits radar or acoustic signals and the reflection of the signal is used to collect traffic data.</p>
<p>Infrared Technology</p>		<p>Vehicles can be observed by means of disturbances in the infrared beam.</p>
<p>Laser Detection</p>		<p>A laser is installed above the roadway, emitting a beam aimed at a photodiode array placed on the pavement. The beam breaks when vehicles pass underneath the laser.</p>
<p>Inductive Loop Detector</p>		<p>A circular loop is placed on the pavement, and connected to an electronics box. Vehicles passing induce a current in the loop, allowing detecting them.</p>

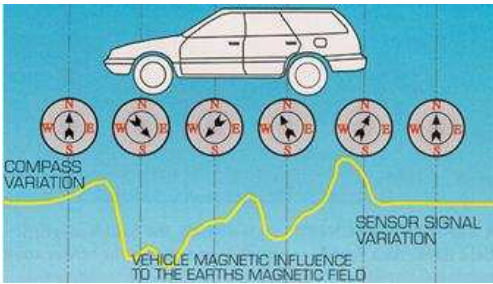
Technology	Image	Brief description
Wireless Magnetic Technology		Magnetic equipment is installed on the road. Vehicles are detected by means of perturbations on the magnetic field.

Table 1. Detector technologies for traffic counts in a traffic network.

Normally, the most common detector used in urban areas are induction loops, due to its low costs and power requirements, although new technologies (such as video or infrared detection) continue to improve and have been successfully implemented.

Most common parameters to be measured in traffic systems implementations are quantities such as traffic volumes, speeds, occupancy, and even weight or length of the vehicles. Depending on the needed data, different technologies of detectors will be installed.

When automated devices are used, this data is typically collected in a continuous way and at a relatively fine resolution, except for communication or technical failures. Other times, when there is not any permanent infrastructure in the area, it is necessary to turn to provisional TCS's installations or even traffic counts carried out by qualified persons. As seen at the first part of this chapter, it is necessary that every data fulfils the flow conservation laws all along the network, and that is why several methods of adjustment and optimization are developed to obtain a reliable set of data to analyse the transport network.

Generally, two types of errors can be committed during the acquisition of field flow volume data in a road:

Admissible errors. Errors that are within the measuring device's tolerance and, consequently, they depend on the precision defined for each device by the manufacturer. For example, if the manufacturer of the detectors in the TCS indicates 3% reliability, it means that if one of the measurements is x^{obs}

= 924, the real value $x^* \in [924(1 - 0.03), 924(1 + 0.03)]$. In practice, the admissible range of error is often higher, since margins tend to increase with use and over time.

Inadmissible errors. These are errors that not only give erroneous information, but also invalidate the work done. They can be due to detector malfunctioning (e.g., failure to record passing vehicles, constant recording of non-existent vehicles, always counting an arbitrary number, etc.) or to failure on the part of the person who handles the detector (e.g., failure to set the counter to zero, erroneous readings, even human error in the installation, reading or recording of the data, etc.).

Accordingly to this, apart from possible TCS's errors due to several reasons inherent to traffic flows (admissible errors), it is possible that an undetermined number of data come from loop detectors that are malfunctioning and reporting erroneous data to the Traffic Management Centre (inadmissible errors).

Therefore, when an analyst is creating a traffic model to study an specific problem in a network, he has to deal with a set of data into which he may find several levels of data's accuracy, that would force him to introduce a pre-processing of the data previously to the analysis of the traffic network in order to determine and improve the accuracy of the data.

On the first hand, as seen in section 2.1, all along a network it is necessary that traffic volume data will be consistent and satisfy flow conservation, but this rarely occurs. So, within a set of observed data there are "exact data" and "admissible erroneous data" that may need an adjustment by means of different methods that have been discussed previously.

But, on the other hand, it can be found either a group of missing data or incorrect with inadmissible errors from non-working detectors whose pre-processing will consist on assigning them new values obtained by imputation data techniques. That is an added problem to the analyst of the transport network during the pre-processing of the data, because he will have to deal with a set of data with a high level of uncertainty in many cases.

A graphical representation of this problem can be found in the following figure.

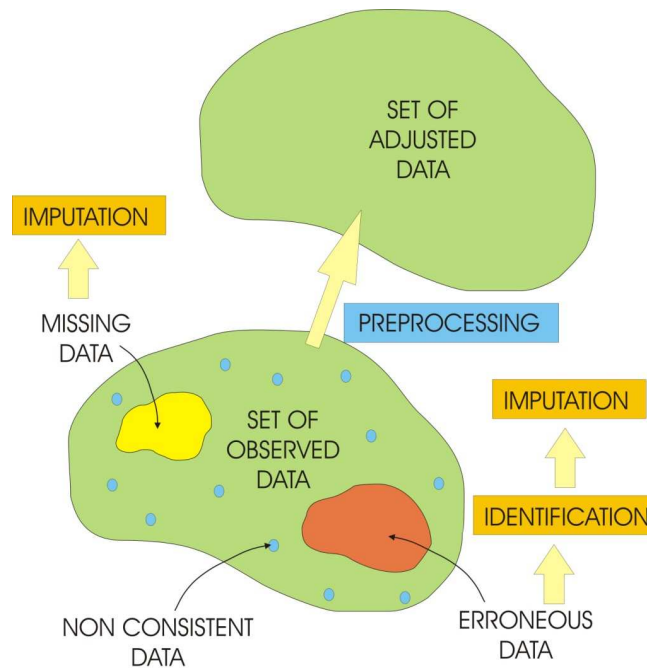


Figure 4. Schema of general pre-processing of observed data in a traffic network.

In the absence of error-checking procedures, this missing and erroneous data might lead to distorted optimization of the data during the processing, so it is required to try to identify the wrong data to pre-process them in a different way from the rest of the data set.

Some researchers analyse different ITS's installed in big cities all along the world to try to estimate the magnitude of the problem of missing and erroneous data in a traffic network. They find that, due to technical failures of TCS's or general measure error of the devices, it is very common to have an amount of missing data.

Turner et al. (2000) make an analysis of recorded historical data from several cities in Unites States and found that there were almost a 25% of missing data or under suspicion to be erroneous. Along the same lines, Nguyen and Scherer (2003) indicate that in Virginia Department of Transportation at least the 25% of the detectors are working offline or have any kind of failure at any given time and, moreover, even properly working detectors often have until a 5% of missing data, as suggested by Kwon et al. (2004).

2.2.1. The problem of missing data: Imputation data.

The problem of lack of information due to missing field flow data is very often in transportation studies, so multiple methods have been developed for the estimation of new correct values for these data.

Most of methods for imputation data involve the estimation of missing or suspicious data using statistical procedures that allow accurate imputation based on current observations. These statistical methods can be applied in different forms that can be summarized in the followings:

- Using overall series: This technique replaces missing values for the statistical mean of observed values in the data set.
- Using a period within the series in observation is missing: If there are missing values that occur in a particular period, they are substituted for observed values within that specified period.
- Using adjacent observations: In this algorithm the analyst specify a “sliding window” size and computes the statistical using observations before and after the interval of missing values.
- Interpolation: These algorithms replace missing values by interpolating from previously observed values. These techniques involve moving averages, exponential smoothing, linear splines, cubic splines, etc.
- Regression Imputation: These models fill in the missing values using predicted values using regression of a given variable on other variables in the analysis
- Time Series: It is used most of all for forecasting techniques to estimate values for one to several intervals into the future.

The easiest methods involve linear regression, using close (spatially and temporally) observed data to estimate the missing value, according to a past set of data.

Chen et al. (2003) and Nguyen and Scherer (2003) recommend estimating a linear regression model based on neighbouring loop detectors by using historical data.

Al-Deek and Chandra (2004) propose estimating a set of linear regression models where every missing data is related to data at a nearby detector, and then they take the median of all estimates generated in this way. The advantages of linear regression models are their simplicity, ease of computation, and ease of interpretation. However, they are not very useful when neighbouring data is missing as well or not available. So, Al-Deek and Chandra's method is more robust for missing data if there are individual neighbouring detectors, but still cannot be used when all neighbouring data is missing (i.e., when all detectors in a particular area are affected by a power failure).

Kwon et al. (2004) suggests a combination of linear regression with non-normal Bayesian imputation. This procedure estimate a linear regression model, as described by Al-Deek and Chandra (2004), together with the deviation between each past observation and the estimate obtained with the linear regression predicted value. Missing data are imputed by performing a linear regression and, later, applying a deviation sampled from the past set.

Other methods remove the requirement of available neighbouring data and use data only from the missing detector to perform the imputation. Nguyen and Scherer (2003) mention that missing data can be replaced by using historical averages, and Gold et al. (2000) suggests replacing missing data of a detector in a period of time by the average of observed data of that particular detector at nearby time periods. They refer this operation by a “factoring-up” approach.

These approaches can be more reliable, as long as they do not depend on neighbouring data that may be not very accurate itself; however, they are less able to represent current conditions if they perform differently from historical norms.

More recently, Zhong and Sharma (2003) introduce improved imputation methods by means of incorporating correction factors and data from both before and after the failure periods into the traditional models. They find that most imputation models from highway agencies only use historical data and the information available from detectors after the failure period is usually neglected. So, imputation techniques may provide more accurate estimates incorporating more data information.

Also, some new techniques based on Artificial Intelligence and on neural networks, in particular, are being developed (Silva-Ramírez, 2007; Tussel, 2002). Other methods based on weighted least squares regression also exist, such as the methods submitted by Kwon et al. (2008). Certain authors (Kaczmarek, 2005; Marzano et al., 2008; Rudy et al., 2008) propose methods based on the characteristics of erroneous traffic data in urban networks, supplemented with the latest data imputation models (Lee et al., 1998; Geng and Wu, 2008).

2.2.2. Detecting erroneous data

The problem of missing data is very common and easy to detect, but also it is necessary to take into account that there may be data that can be incorrect (due to detectors malfunctioning or other causes) and therefore, they must be subject to imputation as well. To identify this erroneous data within a set of data from a network is very difficult and it has been considered by many researchers.

The most common process to identify these data is the comparison with historical data, or even including in the analysis fundamental physical relationships.

Although it is not possible to evaluate every single value of data as either correct or incorrect, observation of general patterns and internal consistency can be used to mark data that are highly unlikely or physically impossible

Zadeh (1996) develop the Continuous Set theory that is based on the construction of a set of decision rules, by phrasing the decision in natural

language, that later would be included in a decision table showing actions to be taken for each combination of states.

The *state* of each data depends on a preliminary classification into several levels of reliability (e.g., Probably correct, Maybe correct, Probably incorrect and Absolutely incorrect), and this classification is made according to three inputs or consistency criterion listed below:

Fundamental consistency observation. Data should be consistent with basic notions of traffic flow theory and should be physically possible. So, in a first revision of the flow in the network two questions have to be formulated: Is it consistent with basic traffic laws? Are the volume and density measurements reasonable?

Network consistency. Data should be related to nearby measurements in space and time. So, the measured data must be compared to upstream and downstream flows to ensure its accuracy in an acceptable range.

Historical consistency. If historical observations at the same location are available, they can provide insight to the plausibility of current observed data. Practice tells us that the values measured on a road are almost always given for an interval. Values outside of the interval may be plausible, but they indicate outliers, an anomaly that should alert the control service. The historical values constitute a basis for determining the boundaries of the interval in which normally consistent values must be found.

Many authors mathematically treat the Continuous Set theory, as Von Altrock (1995). Later, Payne et al. (1976) use fundamental physical relationships in their analysis to mark data with physically impossible values for volume, speed, and density. They identify five kinds of detector errors and suggested several methods to detect them from 20-second and 5-minute volume and occupancy measurements. These methods determine minimum and maximum flow, density, and speed, and according to that, they declare a sample to be wrong if they fail any of the tests.

Turner et al. (2000) and Chen et al. (2003) look at combinations of data that are impossible, in order to identify this erroneous data, such as zero volume and positive occupancy.

As a matter of fact, Chen et al. (2003) present algorithms to detect bad loop detectors from their outputs, and make a missing data imputation from neighbouring good loops. They find that there is much more information in how detectors behave over time, because empirical observations show that good detectors behave very differently over time from bad detectors. So their algorithm makes diagnoses based on the sequence of measurements from each detector over a whole day.

More sophisticated methods are developed by other authors, such as defining an acceptable set of volume/density values (Nihan et al., 1990), data storage rates (Nihan et al., 2002), or statistical entropy (Al-Deek and Chandra, 2004).

On the other hand, Coifman (1999) and Vanajakshi and Rilett (2004) mark suspicious data taking into account observations from nearby detectors.

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

So far, this study has dealt with problems involving vehicle counts in traffic networks, but in this third section of the thesis, the main subject is slightly different. It is also related to the O/D matrix obtainment, but this time, the object of the study deals with passengers' counts in urban public transport.

When planning public transport networks, it is crucial to know the real 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 is 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. (2011). Smart cards are very similar to credit cards and they are used instead of traditional systems such as paper tickets or magnetic stripe cards. Each smart card is identified by a unique serial number and can be registered to a named person, or they can be anonymous. These smart cards are able to retain information such as the amount of credit left needed for payment on every load station.

So, the big potential of the generalized use of these smart cards is the possibility of recording lots of data about passengers' behaviour that, eventually, may serve to achieve a better design of public transport networks and vehicles planning optimization (Bagchi and White, 2005).

So, through smart card systems, transport service providers may have access to:

- Larger volumes of personal travel data
- Have the possibility to link those data to the individual card and/or traveller
- Have access to continuous trip data during longer periods of time than it is possible to obtain using existing transport data sources
- Identify the kind of most frequent customers

Smart cards improve the quality of data (Dempsey, 2008) 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.

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 are adopted in certain exceptional cases (Qing et al., 2009) (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 transport services at a global level, and the fact that passenger tickets need to be scanned twice means double investment.

In the case of smart cards, there is a wide research by several authors trying to estimate the alighting point in order to obtain the O/D matrix. Normally they assume that the next transaction occurs after alighting, because the system of smart cards has boarding validation only.

Barry et al. (2002) develop a method to estimate the alighting station for the New York subway system based on two assumptions:

- After a trip, users will return to the destination of the previous trip station.
- At the end of a day, users will return to the station where they boarded for the first trip of that same day.

Later, Zhao et al. (2007) propose a method to forecast the alighting point for rail boarding transactions in the Chicago CTA system, focusing on rail boardings followed by a bus boarding transaction. They consider rail stations inside a 400 m radius of distance as the alighting station depending on next boarding bus stop. To apply this method, they make the same assumptions as Barry et al. (2002) but also contemplate that the maximum walking distance is 400 m, or 5 min. Trépanier et al. (2007) develop an object-oriented method to estimate the alighting bus stop in the bus system of Gatineau STO. They also use assumptions from Barry et al. (2002) and use the distance to the next

boarding as the principal criteria to determinate the alighting bus stop but add the possibility of considering the next day, even using weekly travel patterns to complete missing information. Zhao et al. (2007) achieve a 71% success rate in estimating alighting stations for rail boardings, while Trépanier et al. (2007) obtain a 66% success for the bus-only Gatineau system.

Munizaga et al. (2010) propose a method to estimate the alighting station in a multimodal public transport system, where boarding transactions are observed in a complex network in which users travel using the Metro and buses and sometimes validate their trip in a bus station instead of doing so directly on a bus. The assumptions they make are:

- Each card corresponds to a user.
- The nearest station to the next boarding bus stop within a 400 m radius or 5 min walking distance is the alighting station.
- In case of the last transaction of the day, as in previous works, is assumed that its destination is close to the point where the first trip of the day began, finishing the daily trip cycle for that particular user.
- If there is only one trip per card, no inference is possible with single day information.

The basic idea is to follow the trip chain of a card and identify the alighting position (bus or Metro station) by looking at the position and time of the next boarding.

Munizaga et al. (2010) apply their method to two 1-week datasets obtained for different time periods. From the data available, they obtain detailed information about the time and position of boarding public transportation and estimate time and position of alighting for over 80% of the boarding transactions.

On the other hand, 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.

3 Chapter 3: MATERIALS AND METHOD

CHAPTER 3 Materials and Method

In view of the theoretical bases detailed above, this section introduces the model created to deal with uncertainty and solve transport problems.

In order to facilitate the explanation of the proposed methodology, a simple example (Figure 5) is presented to demonstrate the method, followed by the data to be used and the inputs required.

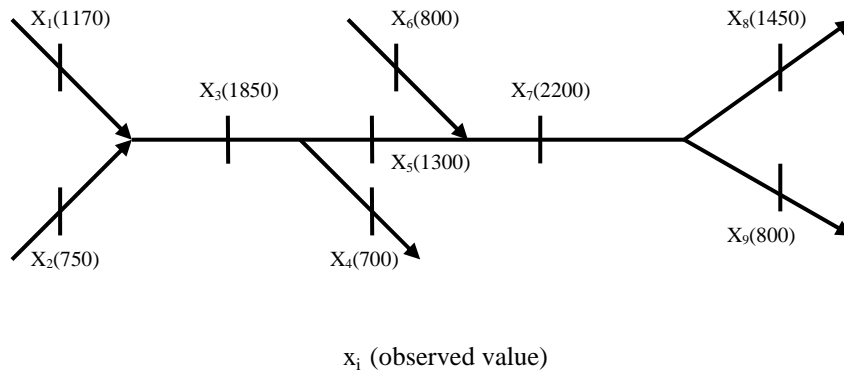


Figure 5. Simple network

Consider the situation shown in Figure 5, in which traffic volumes are observed at the various links, as marked. Theoretically, the total 'incoming volumes' should be equal to the total 'outgoing volumes' at any node in the network so that the law of conservation of flow is satisfied. In real life, this is not usually the case (as in Figure 5), particularly when the network is large. It is known that a large proportion of traffic measurement devices in the field always have some error. Consistency in volume counts in different links is critical to

ensuring the integrity of the results of any Intelligent Transport System (ITS)-related algorithm. The following subsection therefore introduces the different sources of data that may be found in any transport problem.

3.1. Data

Different kinds of information are available:

Fixed numbers: quantitative information in which it is 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 no errors (only in the case of potential failures in the devices) will be considered fixed data. As a consequence of the ticketing systems on a transit line, the boardings will also be considered fixed numbers. The same criterion is followed when no passengers board the vehicle.

Crisp numbers: quantitative numerical data with errors, such as loop detector data or field data, which will have different membership functions depending on their reliability. They are considered to approximate their values, and are allowed to change within a range defined by the parameter α_i . For instance, if a TCS installed on a network counts 50 vehicles, the true number of vehicles may be 47 or 52, etc.

Fuzzy information: qualitative information (from an analyst's viewpoint). We may want to codify subjective measurements, such as congested traffic in a lane; or low, medium, almost full or almost at the limit of road capacity, depending on how the specific problem is defined and whether a maximum, V_{max} is established. For instance, for a vehicle capacity of 50 passengers, 'medium' could correspond to a load ranging from 21 to 29, with a central or most plausible value of 25.

Missing information (when no information is available): in this case, since no information is available, or it was detected as erroneous, and it will not be used as input data, any solution may achieve the same membership grade. This means that every solution is possible. This value will be imputed in light of network consistency and data redundancy.

The first step is to define membership functions to represent the available information. Figure 6 shows the shape of the membership function with the central value x_i^{obs} and a range $[x_i^{obs} - \alpha_i \cdot x_i^{obs}, x_i^{obs} + \alpha_i \cdot x_i^{obs}]$, where α_i is a constant higher than 0. The adjusted value x_i will be found within the range defined by the base of the membership function.

A membership function is convenient for representing the idea that the adjusted value should be an integer ≥ 0 and 'close' to the observed value. Hence, the acceptability of the adjusted value gradually diminishes as it deviates from the observed value. Figure 6 shows the membership function for each kind of information.

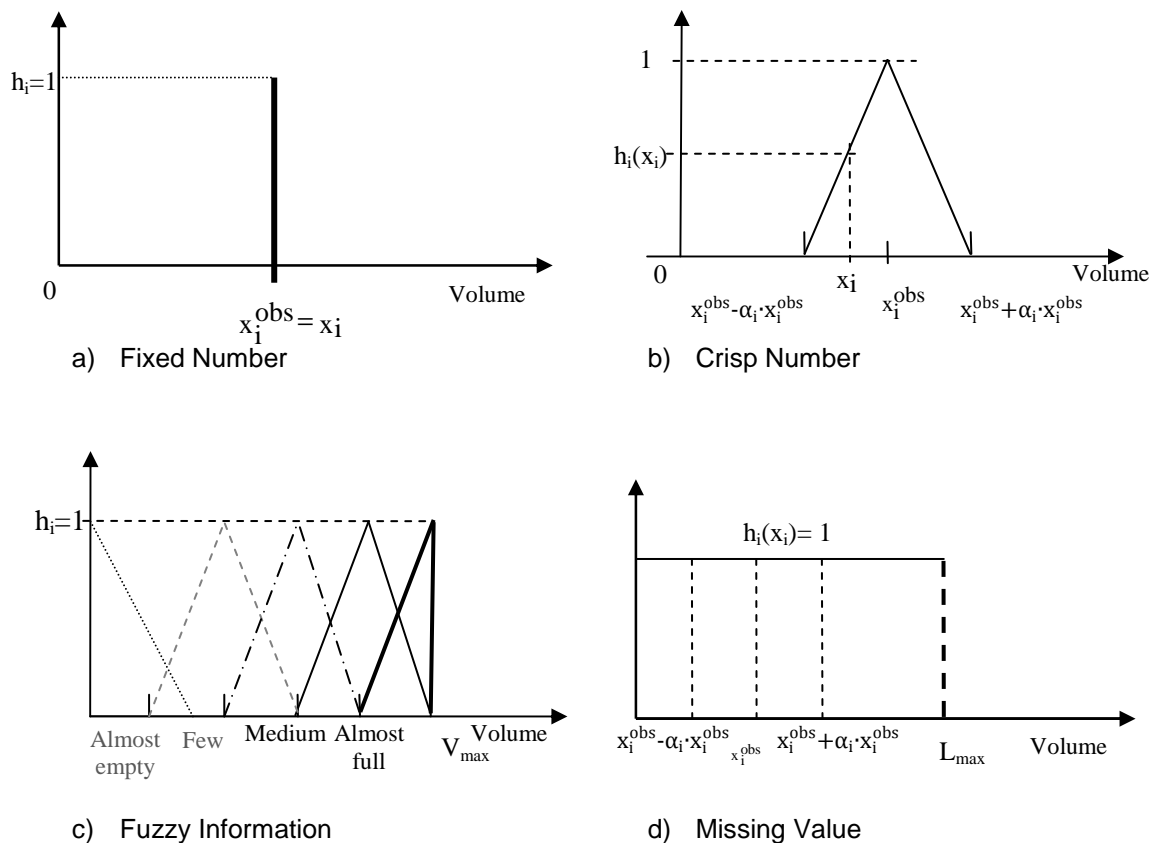


Figure 6. Membership functions for (a) fixed number, (b) crisp number, (c) fuzzy information and (d) missing value

In the case of a fixed number (Figure 6 (a)), we assume that the observed value has no error so it is the same as the adjusted value.

Based on the literature on full fuzzy linear programming using symmetric triangular fuzzy numbers (Lotfi et al., 2009), we assume triangular-shaped membership functions for cases (b) and (c) in Figure 6. This representation is computationally convenient (a linear program can be used). Given an observed value (x_i^{obs}) and its tolerance (α_i) (usually expressed as a percentage of the observed value), Eq. 3 defines the membership function. Notice that if additional information about the character of the observed value is available, the shape of the membership function could be modified (e.g., its triangular shape or the length of the base) but the solution would not change:

$$h_i(x_i) = \max \left\{ 0, 1 - \frac{|x_i - x_i^{\text{obs}}|}{\alpha_i x_i} \right\} \quad \forall i; x_i \in x_i^{\text{obs}} - \alpha_i \cdot x_i^{\text{obs}}, x_i^{\text{obs}} + \alpha_i \cdot x_i^{\text{obs}} \quad (3)$$

where x_i is the adjusted value for the i -th variable.

The selection of the constant α_i depends on the judgement of the analyst with respect to the adjusted value's acceptable deviation from the observed value. It is shown that this value allows the analyst to enter the reliability of each datum (i.e. more reliable data will have a lower value of α_i). If only one value of α_i is used for all data, then the defined range has little effect on the final adjusted values, assuming that it is broad enough for a feasible set of solutions to be found.

Finally, in case (d), where there is no observed value either because it is missing or has an inadmissible error, a membership function is assumed in which all adjusted values are possible and they are all given the same membership grade $h=1$.

3.2. Relationships among data

Regardless of the type of information, all data are closely related and are underlain by a high degree of dependence. These relationships are expressed as constraints:

Equality constraints: Constraints related to the conservation of flow at each control point. They are defined by reviewing the flow pattern at each node in Figure 5 as follows:

$$\begin{aligned} X_1 + X_2 &= X_3 \\ X_3 - X_4 &= X_5 \\ X_5 + X_6 &= X_7 \\ X_8 + X_9 &= X_7 \end{aligned} \quad (4)$$

Inequalities: Constraints related to the membership functions:

$$h_i(x_i) \geq h \text{ for } i=1,2,\dots,k \quad \text{Where } h, x_i \geq 0 \quad (5)$$

which means there are $2k+k$ constraints (where k is the number of control points).

In some situations, as in the case of a transit line, there might exist a limit to x_i , such as the maximum load of the vehicle, V_{max} , which will result in a new inequality condition $x_i \leq V_{max}$.

3.3. Optimization criteria

Given a set of observed values, there is an infinite number of combinations of consistent values, $\vec{x} = (x_i)$, each of which satisfies the aforementioned relationships among the data. The set of these combinations constitutes the feasible region A . For a given combination \vec{x} , the membership grade of each component (x_i) in the corresponding fuzzy set is calculated. Two methods of optimization have been used in the past (Kikuchi and Miljkovic, 1999):

- a. by Maximizing the Minimum (MM) $h_i(x_i)$ for all i :

$$\max_{\vec{x} \in A} \min_i \{h_{x_i^{obs}}(x_i)\} \quad (6)$$

- b. by Maximizing the Sum (MS) of $h_i(x_i)$,

$$\max_{\vec{x} \in A} \left\{ \sum_i h_{x_i^{obs}}(x_i) \right\} \quad (7)$$

In case (a) (the MM method) the lowest membership grade for the combination is recorded. By comparing the lowest membership grades among

all the combinations of traffic volumes, the one that has the highest value is chosen as the best combination of a set of adjusted values.

The problem with the objective function used in this method is that there may be several imputations for the adjusted data that produce the same value for $h = \max(\min(h_i))$ (Silva-Ramírez, 2007). Therefore, they would be the same from the objective function point of view, whereas, in fact, some are better than others. The combination (0.9, 0.9, 0.9, 0.9), for instance, would have the same value as (0.9, 1, 1, 1), whereas the latter is better than the former.

In the MS method (case (b)), for a given combination the membership grade of each adjusted value in the corresponding fuzzy set is calculated. The sum of the membership grades among all the combinations of traffic volumes is recorded, and the one that has the highest sum of membership grades is chosen as the best combination of a set of adjusted values.

This method often leads to some values having $h=0$, despite the fact that almost all the rest are 1, which is of no interest. It has been shown that whereas several of the observed values have $h=1$, other values have lower h and even $h=0$, as can be seen for the third adjusted value in Table 2. This situation is not desirable either, since it allows a set of values with some $h=0$ to be considered, providing the sum is the maximum.

x_i	Observed Value	max-min(h)		max Σh_i	
		Adjusted value	h_i	Adjusted value	h_i
x_1	1,170	1,195	0.7863	1,285	0.0171
x_2	750	772	0.7067	750	1.0000
x_3	1,850	1,967	0.3676	2,035	0.0000
x_4	700	656	0.3714	635	0.0714
x_5	1,400	1 311	0.3643	1,400	1.0000
x_6	800	797	0.9625	800	1.0000
x_7	2,200	2,108	0.5818	2,200	1.0000
x_8	1,450	1,358	0.3655	1,400	0.6552
x_9	800	750	0.3750	800	1.0000
	min h_i		0.3643		0.0000
	sum h_i		4.8811		5.7437

Table 2. Results of MM and MS in the simple example of Figure 5

To solve the aforementioned problems, and with the aim of obtaining the best adjustment data by optimizing all the observed data, a new fuzzy optimization method, Bilevel Optimization (BO), is proposed herein. The method consists of maximizing $\min h_i(x_i)$ for all i at the first level and, after this has been done, applying a second level of optimization by maximizing $(\sum_i(h_i(x_i)))$, considering only those combinations that have $\max(\min h_i(x_i))$.

Thus, the combination with the highest sum is selected from among all the combinations that could maximize the lowest membership grade. The value with the lowest membership grade is taken into consideration, and also all the other observed data.

This method is a two-step or bilevel optimization method:

Step 1. Obtain the set of all the combinations \vec{x} that give the same maximum minimum value h or, in other words, the set that satisfies the following expression (Eq. 6): $\max_{\vec{x} \in A} \min_i \{h_{x_i^{obs}}(x_i)\}$. This set is named A_1 .

Step 2. Obtain the combination $\vec{x} \in A_1$, that satisfies Eq. 7, $\max_{\vec{x} \in A_1} \{\sum_i h_{x_i^{obs}}(x_i)\}$.

Table 3 shows the results of this new model, compared with those of the MM and MS methods.

x_i	Observed Value	max-min(h)		max $\sum h_i$		BO	
		Adjusted value	h_i	Adjusted value	h_i	Adjusted value	h_i
x_1	1,170	1,195	0.7863	1,285	0.0171	1,217	0.5983
x_2	750	772	0.7067	750	1.0000	750	1.0000
x_3	1,850	1,967	0.3676	2,035	0.0000	1,967	0.3676
x_4	700	656	0.3714	635	0.0714	656	0.3714
x_5	1,400	1,311	0.3643	1,400	1.0000	1,311	0.3643
x_6	800	797	0.9625	800	1.0000	847	0.4125
x_7	2,200	2,108	0.5818	2,200	1.0000	2,158	0.8091
x_8	1,450	1,358	0.3655	1,400	0.6552	1,358	0.3655
x_9	800	750	0.3750	800	1.0000	800	1.0000
	min h_i	0.3643		0.0000		0.3643	
	sum h_i	4.8811		5.7437		5.2887	

Table 3. Comparison of results of existing fuzzy optimization methods and BO

Table 3 shows that the BO method obtains the maximum minimum membership grade while optimizing the rest of the adjusted values.

The following are the mathematical steps involved in addressing the optimization problem:

1. Use fuzzy numbers to represent observed values.
2. Formulate the objective and constraints.
3. Solve as a mixed linear integer-programming problem.

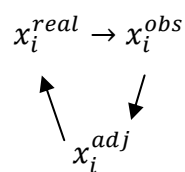
3.4. Results

Certain indicators should be borne in mind when the results are analysed:

- The first indicator is the lowest value of h , which indicates the membership grade of the worst-adjusted value (the degree of compatibility between the adjusted value and the observed value). If the value of h is close to zero, then the adjusted value is close to the right or left end of the base of the membership function; if the value of h is close to one, then the adjusted value is close to the observed value. Therefore, the solution in which the lowest value of h is maximum is chosen as the best solution in terms of this parameter.

- The second indicator is the sum of h_i . The best solution is where the sum of h_i is maximum, because the adjusted values are closer to the observed values and data integrity is better preserved.

In the examples, networks were studied in which the true values were known, so a third indicator should be taken into account in those examples: the average difference between the true consistent values and the adjusted values, which measures the goodness of fit of the adjustment. These results are shown in the following chapters.



When the true values are available, the adjusted values may be compared with them in order to obtain the goodness of fit of the method.

3.5. Applications

Once the model had been generated and testing had shown that it gave a better adjustment than the existing methods and that it could deal with uncertainty and use fuzzy (subjective) information, it was applied to three different transport problems selected from all the possible applications of the model.

First, the model was applied to a road network where loop detectors collect data that do not satisfy the law of conservation of flow. The data were adjusted so as to satisfy this condition and to keep as much integrity as possible. The main conclusions of this application are presented in the first paper published, as described in section 4.2.

Second, the method was applied to detect malfunctioning TCS when no historical data were available. The main conclusions are given in the second published paper (see section 4.3).

The model was also applied to a third transport problem: adjusting boarding and alighting passengers on a public transport route. The major conclusions are given in the third paper published, as described in section 4.4.

Many other applications are sketched out in the final section of this thesis (Future research).

4 Chapter 4: ARTICLES

CHAPTER 4

Articles

4.1. Introduction

The study of any transport system requires enormous quantities of data and an understanding of their dependence on each other. Arguably, for road traffic, the volume is one of the most important traffic datum of them all. Field data is generally inconsistent, and therefore they need to be processed in a way that will make them consistent before they can be used in algorithms for prediction, monitoring and decision-making purposes. The methods used to estimate Origin-Destination (O/D) matrices are based on the hypothetical availability of precise traffic volume data and reliable preliminary O/D data. The input data for most traffic networks, however, are either unavailable or contain measurement errors, as in the case of traffic counts and sensor speed measurements. In the past, certain methods were applied to adjust the observed values so they would comply with flow conservation laws at each network node, aside from other requirements that values need to meet before they can be used as input data in traffic planning algorithms. They are the so-called classic methods. A number of important publications on fuzzy logic have been submitted over the past twenty years, although most of them are based on the fields of deduction and control in situations of complex behavior.

Lost data processing is another frequent issue. When available input data exist, they often contain errors due to the sensors' operating faults. Most efforts have focused on processing 'Missing values', and on detecting and debugging them. Inconsistencies have been avoided by using redundant or related information. Some classical techniques are: imputation by mean, median, regression or hot-deck. Recently, some new techniques based on Artificial Intelligence and, in particular, on neural networks are being developed.

The aim of this thesis is to propose a new method whereby field data could be pre-processed to make them consistent while preserving their integrity as much as possible, and which would include their reliability as perceived subjectively by the analyst. The method is based on fuzzy logic and is intended to optimize the solution obtained. The result would be a reliable solution that comes close to the observed values, thereby resolving measurement errors in Traffic Counts Stations (TCS) and the method is also able to detect which is the most likely TCS to be failing. The method also allows field data to be processed when there are lost values. The results of the aforementioned applications are shown in the first two papers published by the PhD candidate and they are shown in section 4.2, 4.3 and 4.4.

Regarding to transit and public transport, the operation planning and analysis is a concern. Most current ticketing methods can be used to record where passengers get on board but not where they alight. Current methods are unable to make a proper adjustment of boardings and alightings based on the available data unless they do alighting counts, which is very costly. As a spread of the research work developed in this PhD, the proposed fuzzy logic method has been slightly modified and applied to a transit line whereby counts are made at fewer stops and fuzzy information on alightings and/or vehicle loads between consecutive stops are used to make the boarding and alighting adjustment. Fuzzy information can be obtained by the vehicle's driver or an on board observer, which makes it less costly than the counting method. The proposed method presents many benefits: firstly, 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 and; on second hand, 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. The results of this application are shown in the third published paper (see section 4.4).

4.2. Bilevel fuzzy optimization to pre-process traffic data to satisfy the law of flow conservation

De Oña, J., Gomez, P. and Merida-Casermeiro, E. (2011). Bilevel fuzzy optimization to pre-process traffic data to satisfy the law of flow conservation. *Transportation Research Part C*, 19 (1), pp. 29-39. <http://dx.doi:10.1016/j.trc.2010.02.005>

For 2011, the journal TRANSPORTATION RESEARCH PART C- EMERGING TECHNOLOGIES had an Impact Factor of 1.957 and is within Quartile Q1 in the Category Transportation Science & Technology.

Category Name	Total Journals in Category	Journal Rank in Category	Quartile in Category
Transportation Science & Technology	28	5	Q1

The preliminary results of this work were presented at the IV Road Andalusian Meeting in Jaen, held in October 23-26th 2007; at the 87th Annual Meeting of Transportation Research Board in Washington D.C., held in January 13-17th 2008; and at the VIII Transport Engineering Conference (CIT2008) which took place in A Coruña in July 2-4th 2008.

Abstract

Traffic data obtained in the field usually have some errors. For instance, traffic volume data on the various links of a network must be consistent and satisfy flow conservation, but this rarely occurs. This paper presents a method for using fuzzy optimization to adjust observed values so they meet flow conservation equations and any consistency requirements. The novelty lies in the possibility of obtaining the best combination of adjusted values, thereby preserving data integrity as much as possible. The proposed method allows analysts to manage field data reliability by assigning different ranges to each observed value. The paper is divided into two sections: The first section explains the theory through a simple example of a case in which the data is equally reliable and a case in which the observed data comes from more or less reliable sources, and the second one is an actual application of the method in a freeway network in southern Spain where data were available but some data were missing.

Keywords: traffic counts, fuzzy logic, transport planning, optimization, data consistency, subjective analyst knowledge

4.2.1. Introduction

The study of any transport system requires enormous quantities of data and an understanding of their dependence on each other. Arguably, volume is the most important traffic datum of them all. Field data is generally inconsistent, and therefore they need to be processed in a way that will make them consistent before they can be used in algorithms for prediction, monitoring and decision-making purposes. The methods used to estimate Origin-Destination (O/D) matrices are based on the hypothetical availability of precise traffic volume data and reliable preliminary O/D data. The input data for most traffic networks, however, are either unavailable or contain measurement errors, as in the case of traffic counts and sensor speed measurements. In fact, some studies (Zhong et al., 2004) demonstrate that 50% of the Permanent Traffic Counts (PTCs) set up on highways contain lost data, making it difficult to ignore measurement errors when processing data used to plan, design, control and

manage traffic (Sharma et al., 1996). The existence of errors makes data obtained in the field difficult to manage and to analyze.

In the past, certain methods were applied to adjust the observed values so they would comply with flow conservation laws at each network node, aside from other requirements that values need to meet before they can be used as input data in traffic planning algorithms. The methods used were manual value adjustment, least square adjustment and the maximum likelihood method (Kikuchi et al., 2000). Recently, new methods of value adjustment based on fuzzy logic have been developed to preserve data integrity as much as possible. The methods are: fuzzy regression, fuzzy optimization and necessity-interval-regression method (Kikuchi et al., 2000). A number of important publications on fuzzy logic have been submitted over the past twenty years, although most of them are based on the fields of deduction and control in situations of complex behaviour. Papis and Mamdani (1977) were the first to apply fuzzy logic to transport; specifically, to traffic signal controllers.

Lost data processing is another frequent issue. When available input data exist at all, they often contain errors due to the sensors' operating faults (Kwon et al., 2008). From a formal viewpoint, the problem of debugging input data in order to avoid inconsistency and of assigning values to missing data has generally been analyzed by an area of Statistics (Data Editing and Imputation). Most efforts have focused on processing 'Missing values', and on detecting and debugging. Inconsistencies have been avoided by using redundant or related information. Some classical techniques are: imputation by mean, median, regression or hot-deck (Chambers, 2001; Laaksonen, 1999). Recently, some new techniques based on Artificial Intelligence and on neural networks, in particular, are being developed (Silva-Ramírez, 2007; Tussel, 2002). Certain authors (Kaczmarek, 2005; Marzano et al., 2008; Rudy et al., 2008) have submitted methods based on the characteristics of erroneous traffic data in urban networks, supplemented with the latest data imputation models (Lee et al., 1998; Geng and Wu, 2008). Other methods based on weighted least squares regression also exist, such as the methods submitted by Kwon et al. (2008).

The aim of this article is to submit a method whereby field data could be pre-processed to make them consistent while preserving their integrity as much as possible, and which would include their reliability as perceived subjectively by the analyst. The method is based on fuzzy logic and is intended to optimize the solution obtained. The result would be a reliable solution that comes close to the observed values, thereby resolving measurement errors in traffic counts. The method also allows field data to be processed when there are lost values.

4.2.2. Description of the problem

A simple freeway network is used to explain the method. Consider the situation shown in Figure 7, in which real consistent data are available (Table 4, column 2).

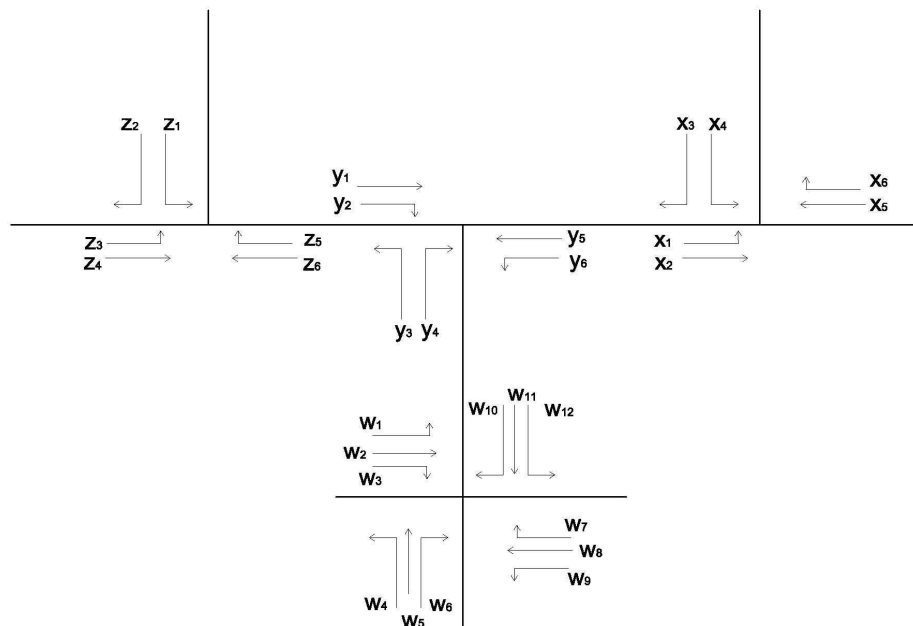


Figure 7. Example of situation in which consistent data are available and are randomized to get traffic counts not consistent to explain the theory

The data are used to simulate a scenario with non consistent data: traffic counts from the database are randomized within $\pm 25\%$ of their values at all intersections to simulate a case in which data is not consistent (Table 4, column 3). Next, the randomly obtained data in the database are considered to be field data; i.e., the observed values (OV).

(1)	(2)		(3)			(4)			(5)			(6)			(7)			(8)			(9)		
	RV	OV	MM		$\alpha=0.4$	MM		α^*	MS		$\alpha=0.4$	MS		α^*	BO		$\alpha=0.4$	BO		α^*			
			AV	h_i	Δ	AV	h_i	Δ	AV	h_i	Δ	AV	h_i	Δ	AV	h_i	Δ	AV	h_i	Δ			
w_1	135	113	128	0.68	7	126	0.63	9	113	1.00	22	113	1.00	22	128	0.68	7	126	0.63	9			
w_2	30	37	39	0.87	9	40	0.91	10	37	1.00	7	37	1.00	7	37	1.00	7	37	1.00	7			
w_3	43	42	42	1.00	1	44	0.92	1	42	1.00	1	42	1.00	1	42	1.00	1	42	1.00	1			
w_4	104	96	95	0.97	9	97	0.97	7	96	1.00	8	96	1.00	8	95	0.97	9	97	0.97	7			
w_5	148	134	152	0.67	4	150	0.61	2	134	1.00	14	134	1.00	14	152	0.67	4	150	0.61	2			
w_6	19	18	21	0.59	2	20	0.82	1	18	1.00	1	18	1.00	1	18	1.00	1	18	1.00	1			
w_7	28	27	30	0.73	2	30	0.64	2	27	1.00	1	27	1.00	1	30	0.73	2	30	0.64	2			
w_8	35	37	40	0.80	5	40	0.91	5	37	1.00	2	37	1.00	2	37	1.00	2	37	1.00	2			
w_9	22	18	21	0.59	1	20	0.82	2	18	1.00	4	18	1.00	4	18	1.00	4	18	1.00	4			
w_{10}	102	78	77	0.97	25	79	0.96	23	78	1.00	24	78	1.00	24	77	0.97	25	79	0.96	23			
w_{11}	175	171	172	0.99	3	169	0.96	6	171	1.00	4	171	1.00	4	172	0.99	3	170	0.98	5			
w_{12}	3	4	4	1.00	1	4	1.00	1	4	1.00	1	4	1.00	1	4	1.00	1	4	1.00	1			
x_1	265	215	253	0.57	12	255	0.71	10	220	0.94	45	220	0.96	45	253	0.57	12	253	0.73	12			
x_2	54	53	62	0.59	8	61	0.77	7	53	1.00	1	53	1.00	1	62	0.59	8	61	0.77	7			
x_3	105	116	109	0.85	4	111	0.93	6	116	1.00	11	116	1.00	11	109	0.85	4	111	0.93	6			
x_4	110	132	130	0.96	20	133	0.99	23	132	1.00	22	132	1.00	22	130	0.96	20	133	0.99	23			
x_5	200	177	168	0.88	32	168	0.92	32	161	0.78	39	161	0.86	39	168	0.88	32	168	0.92	32			
x_6	58	51	48	0.86	10	52	0.97	6	51	1.00	7	51	1.00	7	50	0.95	8	51	1.00	7			
y_1	26	31	26	0.61	0	28	0.79	2	20	0.13	6	20	0.25	6	26	0.61	0	26	0.66	0			
y_2	20	17	15	0.71	5	14	0.62	6	17	1.00	3	17	1.00	3	15	0.71	5	15	0.75	5			
y_3	18	21	21	1.00	3	18	0.70	0	21	1.00	3	21	1.00	3	21	1.00	3	18	0.70	0			
y_4	293	353	289	0.56	4	288	0.61	5	253	0.31	40	253	0.40	40	289	0.56	4	288	0.61	5			
y_5	45	39	39	1.00	6	41	0.89	4	41	0.87	4	41	0.89	4	39	1.00	6	41	0.89	4			
y_6	260	226	238	0.87	22	238	0.89	22	236	0.89	24	236	0.91	24	238	0.87	22	238	0.89	22			
z_1	33	26	29	0.72	4	29	0.75	4	26	1.00	7	26	1.00	7	29	0.72	4	29	0.75	4			
z_2	22	17	20	0.57	2	20	0.87	2	17	1.00	5	17	1.00	5	17	1.00	5	17	1.00	5			
z_3	25	27	29	0.82	4	31	0.92	6	27	1.00	2	27	1.00	2	27	1.00	2	27	1.00	2			
z_4	13	11	12	0.78	1	13	0.61	0	11	1.00	2	11	1.00	2	12	0.78	1	12	0.81	1			
z_5	28	33	32	0.93	4	31	0.87	3	33	1.00	5	33	1.00	5	32	0.93	4	31	0.87	3			
z_6	35	29	28	0.92	7	28	0.93	7	29	1.00	6	29	1.00	6	28	0.92	7	28	0.93	7			
sum h_i			24.04			24.92			27.93			28.26			25.89			25.98					
min h			0.56			0.61			0.13			0.25			0.56			0.61					
Average Δ			7.23			7.13			10.70			10.70			7.10			6.97					

Note: RV (Real Value); OV (Observed Value); AV (Adjusted Value); Δ (difference between RV and AV in absolute value)

* $\alpha=0.65$ for x_i ; $\alpha=0.5$ for y_i and z_i ; $\alpha=0.3$ for w_i

Table 4. Example 1: base data, randomized inconsistent data, adjusted data, and results for different α ranges

Theoretically, in any transport network such as the one shown in Figure 7, the total “incoming volumes” should be equal to the total “outgoing volumes” at any node in the network and in any flow direction in such a way that the law of

conservation of flow is satisfied. In the simulated scenario, (Table 4, column 3), however, it is found that:

$$\begin{aligned}
 x_3+x_5 &\neq y_5+y_6 \\
 x_1+x_2 &\neq y_1+y_4 \\
 y_1+y_2 &\neq z_1+z_4 \\
 y_3+y_5 &\neq z_5+z_6 \\
 y_2+y_6 &\neq w_{10}+w_{11}+w_{12} \\
 y_3+y_4 &\neq w_1+w_5+w_7
 \end{aligned} \tag{8}$$

Actually, this is usually the case, particularly when the network is large. Pentrice (1987) stated that data inconsistency is inevitable even in a well-controlled survey, but volume count consistency at different links is critical to ensuring the integrity of the results of any of the ITS-related algorithms.

When the network becomes larger, the possibility of inconsistency in traffic volume counts increases, so flow conservation is more difficult. The concern in this paper is how to adjust the individual observed volumes to a set of new values that satisfy the flow conservation principle at any point in the network. Furthermore, the adjustment should be such that the integrity of the observed values is preserved as much as possible. To this end, a fuzzy optimization method is used to obtain adjusted values that comply with the law of flow conservation and that resemble consistent real data as closely as possible. In this example, the integrity of the results obtained can be verified with the available real consistent data.

4.2.3. The bilevel fuzzy optimization method

The search for the “best” set of adjusted values is an optimization process that aims to find a set of values close to the observed ones that verifies the conservation of flow principle.

The proposed method is based on the following concept: Each observed value is considered an approximate value represented by a fuzzy number, defined by a membership function. If the value is x , it is interpreted as “approximately x ”. The true value is considered to lie near x . The method attempts to find an adjusted value as close to the observed value as possible while satisfying the conservation of flow at every point in the network. This is accomplished by applying the concept of fuzzy optimization developed in fuzzy set theory.

Given a set of observed values, there are an infinite number of combinations of adjusted values, each of which satisfies the set of flow conservation equations. For a given combination, the membership grade $h_{x_i}(x'_i)$ of each adjusted value (x'_i) in the corresponding fuzzy set (x_i) is calculated. Three methods of optimization could be used:

- a) by Maximizing the Minimum $h_{x_i}(x'_i)$ for all i ,
- b) by Maximizing the Sum of $h_{x_i}(x'_i)$, and,
- c) by maximizing the Minimum $h_{x_i}(x'_i)$ for all i at one level and, after this has been achieved, by applying a second level of optimization by Maximizing the Sum of $h_{x_i}(x'_i)$. Thus, the combination with the highest sum is selected from among all the combinations that could maximize the lowest membership grade. The value with the least membership grade is taken into consideration, and also all the other observed data.

In case (a) (MM method), the lowest membership grade for the combination is recorded. By comparing the lowest membership grades among all the combinations of traffic volumes, the one that has the highest value is chosen as the best combination of a set of adjusted values. This method was already introduced by Kikuchi and Miljkovic (1999).

On the other hand, in the objective function sum of $h_{x_i}(x'_i)$'s (case (b)) (MS method) for a given combination, the membership grade of each adjusted value

in the corresponding fuzzy set is calculated. The sum of the membership grades among all the combinations of traffic volumes is recorded, and the one that has the highest sum of membership grades is chosen as the best combination of a set of adjusted values.

The third possibility is a two step way of optimization or Bilevel Optimization method (BO method). In step one, case a), the lowest membership grade is maximized. In step two, the membership grades that would produce the largest possible $\max(\min(h_i))$ and that would seek to increase the value of all of the h_i at the same time (which would achieve the sum of both) are summed up and maximized.

The MM method can attend to a set of data which its minimum membership grade is maximized but the problem is that an infinite number of combinations could satisfy this condition and the MM method randomly chooses one of them. The BO method chooses a set that while it satisfies that condition; it optimizes the rest of the values, maximizing the membership grade of all the data, so the BO method uses both ways of optimization in order to improve the solution.

The mathematical steps involved in addressing the optimization problem are:

- Use fuzzy numbers to represent observed values
- Formulate the objective and constraints
- Solve as a mixed linear programming problem

The process is explained step by step by using the simple highway network shown in Figure 7.

Using fuzzy numbers to represent observed values

The observed values are “fuzzified” and are considered a fuzzy set with a triangular membership function.

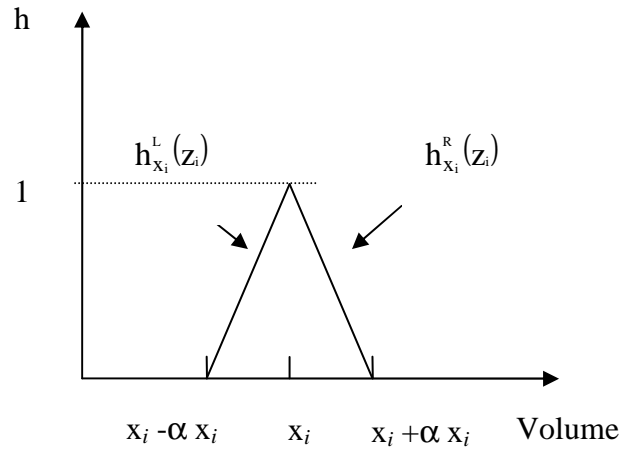


Figure 8. Triangular membership function

Figure 8 shows the shape of the membership function with the centre value x_i and a range $[x_i - \alpha x_i, x_i + \alpha x_i]$, where α is a constant higher than 0. The triangular membership function is not a prerequisite but, in the absence of any other information, this is a reasonable assumption, and such assumption is often used in fuzzy set theory (Zimmermann, 2001).

The selection of the constant α depends on the judgement of the analyst with respect to the adjusted value's acceptable deviation from the observed value. This value allows the analyst to enter the reliability of each datum (i.e. the more reliable data will have a lower value of α than if they were less reliable). If only one value of α is used for all data, the scope of the range has little effect on the final adjusted values, once it is broad enough for a feasible set of solutions to be found.

The membership function is defined for the left- and right-hand sides of the triangle. For an observed value of x_i and the assumed range $[x_i - \alpha x_i, x_i + \alpha x_i]$, the general expression of the membership functions is:

$$h_{x_i}(x'_i) = \begin{cases} h_{x_i}^L = \frac{x'_i - (x_i - \alpha x_i)}{\alpha x_i} & \text{if } x_i - \alpha x_i < x'_i \leq x_i \\ h_{x_i}^R = \frac{x'_i - (x_i + \alpha x_i)}{-\alpha x_i} & \text{if } x_i < x'_i \leq x_i + \alpha x_i \end{cases} \quad (9)$$

In this formula $-\infty < x_i - \alpha x_i \leq x_i \leq x_i + \alpha x_i < \infty$, the triangular fuzzy number x_i is presented by $(x_i - \alpha x_i, x_i, x_i + \alpha x_i)$.

For the sake of simplicity, a symmetric triangle is used in this paper for the membership function. However, the left and right-hand limits can be set separately. To solve this example problem, it is assumed that the value x_i is the observed value and that the value of $\alpha > 0$. So the value of α is the spread of triangular fuzzy number x_i . The narrower the spread area, the less fuzzy the evaluation data will be, hence more precise. To the contrary, fuzziness is higher and thus more vague and ambiguous when the spread area increases (Tzeu-Chen Han, 2008).

Some authors have researched calibration of the membership function extensively. The classical approach to calibration has been the intuitive trial and error process, in which the analyst modifies the shapes of the membership functions little by little until the predicted output approximately fits the output data obtained from the real world (Chakroborty and Kikuchi, 2003). However, this process is time consuming. Other authors have developed a systematic way of carrying out the trial and error process (Wang and Mendel, 1992a, 1992b, 1992c; Homaifar and McCormick, 1995). The purpose of calibration is to modify the membership functions of the Fuzzy Inference System (FIS) so that the outcome predicted by the model is equal (or nearly equal) to the outcome obtained in the real world. Therefore, Chakroborty and Kikuchi (2003) presented a method in which a representation framework allows the FIS parameters to be modified in relation to the bases. FIS outputs are dictated by the parameters that define the membership functions of the fuzzy sets appearing in the antecedents and the consequents of the rules and the algebraic operators used for the logical connectives and to determine the final inferred value. They have developed a procedure that calibrates the membership function of the fuzzy sets by transforming the inference system into an Artificial Neural Network format. They have applied this procedure to the complex control task of car-following, but this procedure has not been applied yet to an urban transport system or a large-scale civil infrastructure system.

Formulating the objective function and its constraints

In a fuzzy number representation of observed values, fuzzy optimization techniques would be used to search for the adjusted values. The mathematical

formulation of the three proposed methods used to solve the problem would be as follows:

A. MM method:

$$\text{Max}(h) \text{ where } h \text{ is } \min(h_i) \quad (10)$$

Subject to

Constraints related to the membership functions:

$$h_{x_i}^L(x'_i) \geq h \quad h_{x_i}^R(x'_i) \geq h \quad h_i \geq h \quad \text{for } i=1, k \quad (11)$$

which means there are $2k+k$ constraints (where k is the number of control points)

Constraints related to the conservation of flow at each control point. The constraints are defined by reviewing the flow pattern at each node in Figure 7 as follows:

$$\begin{aligned} x'_3 + x'_5 &= y'_5 + y'_6 \\ x'_1 + x'_2 &= y'_1 + y'_4 \\ y'_1 + y'_2 &= z'_1 + z'_4 \\ y'_3 + y'_5 &= z'_5 + z'_6 \\ y'_2 + y'_6 &= w'_{10} + w'_{11} + w'_{12} \\ y'_3 + y'_4 &= w'_1 + w'_5 + w'_7 \\ x'_i, y'_i, z'_i, w'_i &\geq 0 \text{ for all } i \end{aligned} \quad (12)$$

Where

x'_i, y'_i, z'_i, w'_i integer unknown adjusted values

x_i, y_i, z_i, w_i fuzzy set corresponding to the observed value x_i

$h_{x_i}(x'_i)$ membership grade of x'_i in the fuzzy set x_i , the same treatment for y_i, z_i and w_i ; h is an operational parameter that represents the smallest membership grade among all $h_{x_i}(x'_i)$'s. Where $h_{x_i}^L(x'_i) \geq h$ and $h_{x_i}^R(x'_i) \geq h$,

respectively, show the expressions for the left- and right-hand sides of the triangle.

B. MS method:

$$\text{Max}(g) \text{ where } g \text{ is } \text{sum}(h_i) \quad (13)$$

Subject to the same constraints as in MM method related to the membership functions (Eq. 11) and related to the conservation of flow at each control point (Eq. 12).

C. BO method:

Step 1: The problem is solved using MM method (Eq. 10), and we obtain a value of $h=h^*$.

Step 2: The problem is solved using MS method (Eq. 13) subject to the same constraints related to the conservation of flow at each control point (Eq. 12) as in the MM or MS method, and to the following constraints related to the membership functions:

$$h_{x_i}^L(x'_i) \geq h^* \quad h_{x_i}^R(x'_i) \geq h^* \quad h_i \geq h^* \quad \text{for } i=1, k \quad (14)$$

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

If only Max(h) is performed (case A), there may be several imputations for the observed data that produce the same value for h (Tussel, 2002; Silva-Ramírez, 2007). Therefore, they would be the same from the objective function point of view, whereas, in fact, some are better than others. The combination (0.9, 0.9, 0.9), for instance, would have the same value as (0.9, 1, 1), whereas the latter is better than the former. On the other hand, if the objective function were just Max(g) (case B), some values would show a $h_i=0.00$, despite the fact that almost all the rest are 1.00, which is of no interest. The bilevel optimization process (case C) allows the combination where the remaining membership degrees are the highest ones to be chosen from among all the combinations where the lowest value of h is maximized.

Solving as a mixed linear programming problem

Since every x'_i must be an integer number and h_i are real numbers, this is a mixed linear programming formulation. A mixed linear programming algorithm is formulated for the problem to maximize the membership grade of the adjusted values.

In Figure 7, the mixed linear programming algorithm consists of 90 (3x30 observed volumes) inequality constraints related to membership functions and six equations related to flow conservation.

Introduction of data reliability

The selection of the value of α depends on the judgement of the analyst with respect to the adjusted value's acceptable deviation from the observed value.

In a complex transport network, there may be permanent traffic count stations where count data are fairly reliable, and other nodes where counting is sporadic, as well as points where traffic volumes have not been measured. Therefore, to define the α parameter coherently, the method must allow the analyst to assign different values to the α parameter in order to define the membership functions of each observed value. The values will depend on whether the parameter belongs to a set of data that are highly reliable (permanent traffic count station), averagely reliable (sporadic count) or highly unreliable (lost data).

4.2.4. Example network

As shown in Figure 7, the example consists in analysing a network of 4 intersections, of which three have 6 movements and one has twelve.

In this example, the real consistent data are known (RV) (Table 4 column 2). The data are used to simulate a scenario with non consistent data. The simulated data are considered the OV (Table 4 column 3).

In this example, it is considered that traffic count station W is a permanent station, so the values have maximum reliability and their α parameter is the

lowest, $\alpha=0.3$. The reliability of stations Y and Z is lower so α takes a value of 0.5 (sporadic count stations) and, finally, the data from traffic count station X is supposed to be the least reliable one, so α is assigned a value of 0.65.

4.2.5. Results

In this case, since real data were available, three indicators could be used to verify the goodness of the adjustment of each one of the three optimization methods used (MM, MS and BO methods):

The first indicator is the lowest value of h , which indicates the membership grade of the worst adjusted value (the degree of compatibility between the adjusted value and the observed value). If the value of h is near zero, then the adjusted value is close to the right or left end of the base of the membership function; if the value of h is near 1, then the adjusted value is close to the observed value. Therefore, the solution where the lowest value of h is maximum is chosen as the best solution from the point of view of this parameter.

The second indicator is the sum of h_i . The best solution is where the sum of h_i is maximum, because the adjusted values are closer to the observed values and integrity is more preserved.

The last indicator, for which the consistent real data are available, is the average of the differences between the real consistent values and the adjusted values.

The results for the three methods are given in Table 4, where the adjusted values (AV) and the value of the membership grade (h) for each observed value are shown. The membership grade of the individual AV is computed by entering the adjusted value (x'_i) in the respective membership function $h_{x_i}(x'_i)$. The table also shows the effect of using different α values, depending on the reliability of the observed volumes at each intersection.

Column 1 of Table 4 shows each movement in nodes W, X, Y and Z. Column 2 shows the consistent RV used to obtain the OV that show inconsistencies by randomizing the values within $\pm 25\%$.

Columns 4, 6 and 8 in Table 4 shows the AV, the corresponding values of h (h_i) and the difference (Δ) between RV and AV in absolute value, using the MM method, MS method and BO method respectively for an α parameter of 0.4 in all cases:

MM method's results are shown in column 4. The lowest value of h in column 4 ($h=0.56$) indicates the membership grade of the worst adjusted value. In this case Σh_i is 24.04.

In column 6, MS method's results show that whereas most of the adjusted values get $h=1.00$, other values show lower h and h could even be 0.00, in order to manage the highest Σh_i . The lowest value of h is reached for y_1 ($h=0.13$). This situation, therefore, is not desirable either, since it allows a set of values with some h very close to 0.00 to be considered, providing the sum is the maximum. In this case, Σh_i is 27.93.

The BO method's results are shown in column 8. If columns 4 and 8 are compared, it can be seen that the minimum value of h remains the same ($h=0.56$). However, there has been an increase in Σh_i , which has gone from 24.04 (MM method) to 25.89 (BO method). Thus, this new method allows a combination where the remaining membership degrees are the highest ones to be chosen from among all the combinations with the lowest value of h .

As explained above, introducing the analyst's knowledge of the different precisions of the data he is working with improves the results of the adjustment. This is shown in columns 5, 7 and 9 in Table 4 where the AV, h_i and Δ are calculated, using the three methods for different α parameters depending on the reliability of the data. The α values used in this example have been 0.65 for "X", 0.5 for "Y" and "Z" and 0.3 for "W".

As in the case of the same α for every observed value, for any α parameter, the MM and the BO methods obtain the same and a higher value of h minimum ($h=0.61$) than the MS method ($h=0.25$). However, the latter method obtains a higher value of Σh_i (28.26 versus 24.92 for the MM method and 25.98 for the BO method).

The results shown in column 5 are better than those shown in column 4. This is because the minimum value of h and $\sum h_i$ were higher and the value of average Δ was lower. Similar results are obtained by comparing columns 6-7 and 8-9 for the MS and the BO methods. This confirms the advantage of distinguishing between reliable data and less reliable data or, in other words, of introducing the subjective perception of the analyst.

The last row in Table 4 shows the average of Δ for each of the three methods used. It can be seen that the lowest value (6.97) is obtained for the BO method with different values of α , in comparison to the values of the MM method (7.13) and the MS method (10.70) with different values of α . This shows that the AV obtained with the BO method are closer to the real values than with the other two methods, so this is the method that best preserves the integrity of data.

Real intersections in Andalucía motorway's network

Next, the three methods are used to adjust the traffic volumes of a series of adjacent intersections in Andalusia's freeway network (see Figures 9 and 10) for which real and therefore inconsistent data are available. In this example, the parameter Δ is omitted, and only two parameters have been used to verify the goodness of the adjustment: the lowest value of h and the sum of h_i .

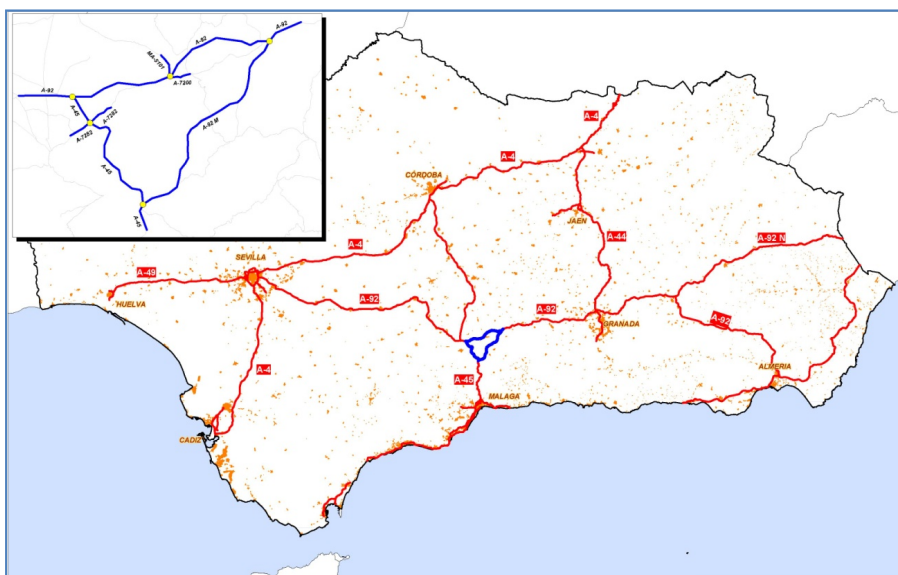


Figure 9. Not consistent real base data set of traffic counts for intersection in Andalusia (South of Spain)

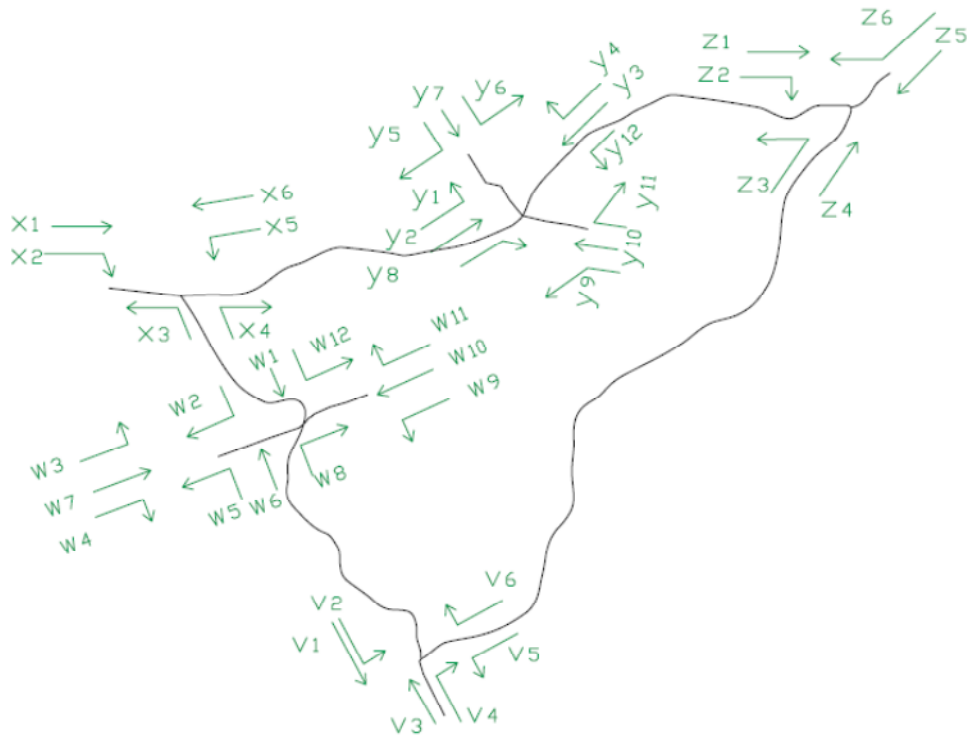


Figure 10. Movements in every node of the network

The network has five intersections, of which three have six movements (intersections V, X and Z), while the other two have twelve potential movements (intersections W and Y). Data is available for all potential movements except for movements v_1 , v_2 , v_3 , v_4 , y_6 , y_7 , and y_8 , whose values were lost. A special membership function with $h=1.00$ always ($\alpha \rightarrow \infty$) was assigned to the lost values so that any adjusted value that met the boundary conditions would always have a membership grade of 1.00 (Figure 11). Table 5 shows that for movements v_1 , v_2 , v_3 , v_4 , y_6 , y_7 and y_8 , the value of h associated to the AV is always 1.00 for the three methods studied and for the hypothesis of equal or different α .

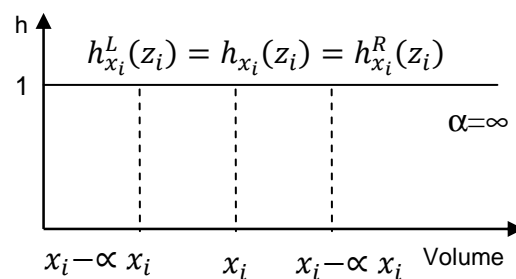


Figure 11. Missing values' membership function

(1)	(2)	(3)		(4)		(5)		(6)		(7)		(8)	
	OV	MM $\alpha=0.1$		MM α^*		MS $\alpha=0.1$		MS α^*		BO $\alpha=0.1$		BO α^*	
		AV	h_i	AV	h_i	AV	h_i	AV	h_i	AV	h_i	AV	h_i
$v_1^{\#}$	-	11,091	1.00	10,819	1.00	10,703	1.00	10,704	1.00	10,951	1.00	10,893	1.00
$v_2^{\#}$	-	1,764	1.00	1,758	1.00	1,743	1.00	1,743	1.00	1,497	1.00	1,555	1.00
$v_3^{\#}$	-	11,085	1.00	11,014	1.00	11,240	1.00	11,240	1.00	10,994	1.00	10,847	1.00
$v_4^{\#}$	-	10,288	1.00	10,307	1.00	10,329	1.00	10,329	1.00	10,575	1.00	10,517	1.00
$v_5^{\#}$	10,865	10,727	0.87	10,809	0.95	10,865	1.00	10,865	1.00	10,618	0.77	10,865	1.00
v_6	1,207	1,174	0.73	1,201	0.95	1,207	1.00	1,207	1.00	1,207	1.00	1,207	1.00
w_1	5,427	5,669	0.56	5,491	0.94	5,427	1.00	5,427	1.00	5,445	0.97	5,445	0.98
w_2	2,714	2,905	0.30	2,719	0.99	2,714	1.00	2,714	1.00	2,714	1.00	2,714	1.00
w_3	3,135	2,914	0.30	3,023	0.82	3,135	1.00	3,135	1.00	3,135	1.00	3,135	1.00
w_4	3,123	3,258	0.57	3,179	0.91	3,123	1.00	3,124	1.00	3,123	1.00	3,123	1.00
w_5	3,735	3,764	0.92	3,773	0.95	3,735	1.00	3,735	1.00	3,735	1.00	3,735	1.00
w_6	5,601	5,313	0.49	5,311	0.74	5,600	1.00	5,600	1.00	5,354	0.56	5,207	0.65
w_7	695	744	0.30	695	1.00	695	1.00	695	1.00	695	1.00	695	1.00
w_8	3,112	3,182	0.78	3,131	0.97	3,112	1.00	3,112	1.00	3,112	1.00	3,112	1.00
w_9	3,880	3,928	0.88	3,907	0.97	3,896	0.96	3,896	0.98	3,880	1.00	3,880	1.00
w_{10}	505	470	0.31	502	0.97	505	1.00	505	1.00	505	1.00	505	1.00
w_{11}	310	289	0.32	307	0.95	310	1.00	310	1.00	310	1.00	310	1.00
w_{12}	904	841	0.30	913	0.97	904	1.00	904	1.00	904	1.00	904	1.00
x_1	4,935	4,588	0.30	4,711	0.54	4,812	0.75	4,812	0.75	4,800	0.73	4,709	0.54
x_2	4,725	5,021	0.38	4,788	0.87	4,848	0.74	4,848	0.74	4,719	0.99	4,715	0.98
x_3	7,236	6,739	0.31	6,905	0.54	7,236	1.00	7,236	1.00	6,990	0.66	6,905	0.54
x_4	1,809	1,777	0.82	1,736	0.60	1,809	1.00	1,809	1.00	1,809	1.00	1,747	0.66
x_5	4,197	4,394	0.53	4,335	0.67	4,197	1.00	4,197	1.00	4,344	0.65	4,348	0.64
x_6	3,350	3,306	0.87	3,197	0.54	3,350	1.00	3,350	1.00	3,351	1.00	3,197	0.54
y_1	1,230	1,176	0.56	1,236	0.97	1,230	1.00	1,230	1.00	1,230	1.00	1,230	1.00
y_2	3,700	3,662	0.90	3,717	0.95	3,700	1.00	3,700	1.00	3,700	1.00	3,700	1.00
y_3	4,255	4,555	0.29	4,307	0.88	4,257	1.00	4,257	1.00	4,555	0.29	4,255	1.00
y_4	1,410	1,509	0.30	1,441	0.78	1,410	1.00	1,410	1.00	1,509	0.30	1,410	1.00
y_5	2,140	2,076	0.70	2,094	0.79	2,140	1.00	2,140	1.00	1,990	0.30	2,140	1.00
$y_6^{\#}$	-	2,320	1.00	2,332	1.00	2,369	1.00	2,369	1.00	2,369	1.00	2,369	1.00
$y_7^{\#}$	-	658	1.00	611	1.00	521	1.00	521	1.00	671	1.00	521	1.00
$y_8^{\#}$	-	1,527	1.00	1,494	1.00	1,691	1.00	1,691	1.00	1,679	1.00	1,526	1.00
y_9	1,150	1,069	0.30	1,131	0.83	1,150	1.00	1,150	1.00	1,150	1.00	1,150	1.00
y_{10}	310	289	0.32	324	0.65	310	1.00	310	1.00	310	1.00	310	1.00
y_{11}	1,013	1,044	0.69	1,023	0.91	1,013	1.00	1,013	1.00	1,013	1.00	1,013	1.00
y_{12}	1,410	1,509	0.30	1,442	0.77	1,410	1.00	1,410	1.00	1,509	0.30	1,410	1.00
z_1	4,960	4,975	0.97	4,968	0.99	4,962	1.00	4,962	1.00	4,962	1.00	4,962	1.00
z_2	2,120	2,051	0.67	2,104	0.97	2,120	1.00	2,120	1.00	2,120	1.00	2,120	1.00
z_3	1,207	1,122	0.30	1,092	0.68	1,207	1.00	1,207	1.00	1,122	0.30	1,087	0.67
z_4	10,865	10,930	0.94	10,973	0.97	10,865	1.00	10,865	1.00	10,950	0.92	10,985	0.96
z_5	9,660	9,850	0.80	9,906	0.92	9,952	0.70	9,952	0.90	9,705	0.95	9,952	0.90
z_6	6,940	6,451	0.30	6,098	0.60	5,870	0.00	5,870	0.49	6,451	0.30	5,988	0.54
	sum h_i	26.16		36.50		40.14		40.85		35.97		38.61	
	min h	0.29		0.54		0.00		0.49		0.29		0.54	

Note: OV (Observed Value); AV (Adjusted Value); * $\alpha=0.2$ for w_i ; $\alpha=0.3$ for z_i and z_i ; $\alpha=0.1$ for rest of cases; # missing values

Table 5. Real intersection in the South of Spain: real base data with missing values, adjusted data, and results for different α ranges

Columns 3, 5 and 7 in Table 5 show the AV and h_i using the three methods for $\alpha=0.1$.

On the other hand, columns 4, 6 and 8 in Table 5 show the AV and h_i using the three methods for different α parameters depending on the reliability of the data. The α values used in this example were 0.2 for “W”, 0.3 for “Z”, and 0.1 for the rest.

As in the previous example, for any α parameter, the MM and the BO methods obtain the same and a higher value of minimum h ($h=0.29$) than the MS method ($h=0.00$). However, MS method obtains a higher value of Σh_i (40.14 versus 26.16 for the MM method and 35.97 for the BO method). Thus, the results demonstrate that the BO method, while keeping the highest minimum of h , attains the best sum of h_i , so the best solution is chosen from among all the possibilities that satisfy the condition of maximizing the minimum h . Furthermore, introducing the analyst’s knowledge of the different precisions of the data he is working with improves the results of the adjustment.

4.2.6. Summary and conclusions

The consistency of the observed traffic data is a concern because in nearly all cases traffic data contain some errors. The degree to which consistency must be satisfied depends on the purpose of the analysis. Processing observed data for consistency is crucial in an analysis where data interrelationships are important.

This paper proposes another step forward in using fuzzy logic optimization to obtain adjusted values. Two examples are given to present and explain the theoretical formulation and computational procedure. The proposed approach is robust enough to deal with other typical data discrepancies in transport situations. It preserves the integrity of observed data as much as possible, and allows the analyst to distinguish between reliable and less reliable data.

The approach is able to:

Preserve the integrity of the observed data as much as possible. There are increasing concerns about data imputation and Base Data Integrity. The principle of Base Data Integrity is an important theme discussed by the American Society for Testing and Materials (ASTM, 1991) and the American Association of State Highway and Transportation Officials (AASHTO, 1992). The principle says that traffic measurements must be retained without modification and adjustment. Missing values should not be imputed in the base data. However, this does not prohibit imputing data at the analysis stage. In some cases, traffic counts with missing values could be the only data available for certain purposes and data imputation is necessary for further analysis. In accordance with the principle of Truth-in Data, AASHTO Guidelines (AASHTO, 1992) also recommends highway agencies to document the procedures for editing traffic data. For traffic counts with missing values, highway agencies usually either retake the counts or estimate the missing values. Estimating missing values is known as data imputation.

Ensure flow consistency at any point in the network; the final estimate satisfies the law of flow conservation.

Handle a large complicated network of any size and shape. The aim is to be able to solve any real problem, as shown in example 2.

Handle data reliability; traffic-responsive control systems require reliable real-time information on the prevailing traffic counts to make sensible control decisions. This requisite is met by using the α parameter to define a different range for the membership function associated to each observed value.

Limit the adjusted value within a tolerable deviation from the observed value, but allowing one tolerance for each value to be defined; this is achieved by using fuzzy logic and the definition of the α parameter.

Be solved in a short computation time. The triangular membership function allows solving the problem using mixed linear programming.

The method is flexible so that it can handle cases in which data are questionable, some of the observed values are known and fixed ($\alpha=0$), and

there are considerable discrepancies in the observed data. The base of the membership function within which a feasible set of solutions is searched should be established according to the acceptable difference between adjusted and observed values.

Finally, the method is applicable to many other transportation problems in which consistency is important.

4.2.7. Acknowledgements

The authors appreciate the reviewers' comments and effort in order to improve the paper.

4.3. Method to Detect Malfunctioning Traffic Count Stations

De Oña, J., Gomez, P. and Merida-Casermeiro, E. (2012). Method to Detect Malfunctioning Traffic Count Stations. IET Intelligent Transport Systems, 6(4):364. <http://dx.doi: 10.1049/iet-its.2011.0102>.

For 2012, the journal IET Intelligent Transport Systems had an Impact Factor of 0.959 and Quartile Q3 in the Category Transportation Science & Technology.

Category Name	Total Journals in Category	Journal Rank in Category	Quartile in Category
Transportation Science&Technology	30	15	Q3

Abstract:

This study presents a method for the automatic detection of malfunctioning traffic count stations (TCS) in a transport system. First, double linear optimisation is used to detect inadmissible errors in the recordings of a series of TCS and next, the TCS that are most likely to be failing are identified. The method has been applied to an urban traffic network showing success rates up to 93% in identifying the TCS that are failing.

Keywords: traffic count errors, linear optimization, transport planning, data consistency

4.3.1. Introduction

In traffic operation management and control field, accurate estimates of the density of vehicle flow density in road networks are very important. Information on traffic density may be ascertained from gross counts taken by loop detectors and other detection devices. However, the counts available may be incorrect due to an improper collection process and errors.

When counting the number of vehicles that travel on a road, two types of errors can be committed:

• **Admissible:** In general, admissible errors are the errors that are within the measuring device's tolerance and, therefore, they depend on the precision defined for each device by the manufacturer. For instance, if the manufacturer of the detectors in the traffic counts stations (TCS) indicates 3% reliability, it means that if one of the measurements is $x_{obs} = 784$, the real value $x^* \in [784(1 - 0,03), 784(1 + 0,03)]$. In practice, the admissible boundary of error tends to be somewhat higher, since margins tend to increase with use and over time.

• **Inadmissible:** These are errors that not only give erroneous information, but also invalidate the work done. They can be due to detector malfunctioning (failure to record passing vehicles, constant recording of non-existent vehicles, always counting an arbitrary number, etc.) or to failure on the part of the person who handles the detector (failure to set the counter to zero, erroneous readings, etc.)

In an intersection with two in (x_1 and x_2) and three out movements (x_3 , x_4 and x_5), the principle of flow conservation should verify that:

$$x_1 + x_2 = x_3 + x_4 + x_5 \quad (15)$$

Let the measurements be taken and the following is obtained:

Case 1 $x_1^{obs} = 800$, $x_2^{obs} = 1200$, $x_3^{obs} = 600$, $x_4^{obs} = 700$ and $x_5^{obs} = 740$.

It is found that the above-mentioned condition is not verified, since: $x_1 + x_2 = 2000$, whereas $x_3 + x_4 + x_5 = 2040$. Are the measurements reliable and therefore they can provide relevant information? Or are they indicating that a detector is failing and giving inadmissible measurements? In this case, and assuming that 3% of errors is admissible, we can indicate the existence of a set of values for the measurements that verifies the condition of conservation flow and is within the tolerance range: $x_1^{adj} = 808$, $x_2^{adj} = 1212$, $x_3^{adj} = 594$, $x_4^{adj} = 693$ and $x_5^{adj} = 733$. Therefore, they should be close to the real values.

Case 2 $x_1^{obs} = 800$, $x_2^{obs} = 1200$, $x_3^{obs} = 1600$, $x_4^{obs} = 700$ and $x_5^{obs} = 740$.

It is found that the above condition is not verified either, since: $x_1+x_2 = 2000$, whereas $x_3+x_4+x_5 = 3040$. However, at present no combination of x_i^{adj} values verifies flow conservation and falls within the 3% tolerance range. The inference would be that one of the measurements was erroneous and a detector must be repaired or replaced (unless there was a human error in the installation, reading or recording of the data).

A number of studies (Kikuchi and Miljkovic, (1999); Wall and Dailey (2003); Vanajakshi and Rilett, (2004); de Oña et al., (2011)) attempt to find a solution to Case 1 (admissible errors) to obtain adjusted data that are consistent with flow conservation laws.

For Case 2 (inadmissible errors) , several approaches (Nihan and Davis, (1987a, 1987b, 1989); Tavana and Mahmassani, (2000)) have been attempted to resolve or diminish count errors after they have been detected, but they do not address how they can be detected.

The methods for trying to detect errors may be classified according to the consistency criterion (Lin et al., (2012)):

- Fundamental consistency: data should be consistent with basic notions of traffic theory and should be physically plausible; establishes upper and lower boundaries for traffic values (e.g. negative values and vehicle volumes that exceed the road's capacity cannot be measured).

- Network consistency: data should be related to measurements that are close in space and time. It is based on flow conservation when several connected nodes in a transport network are studied. This is the type of consistency shown in the preceding example.

- Historical consistency: historical observations can provide insight as to the plausibility of current data. Practice tells us that the values measured on a road are almost always given for an interval. Values outside of the interval may be plausible, but they indicate outliers, an anomaly that should alert the control service. The historical values constitute a basis for determining the boundaries of the interval in which normally consistent values must be found.

In current traffic control centres, detecting a malfunctioning count station is pseudo-automated because historical consistency marks the value interval each observation should have. If a measurement is not within that interval, an alarm is triggered, indicating a potential error in one of the TCS.

The problem arises when no historical values are available or when they exist but may indicate measurements as erroneous when they are actually correct. An incident on the network – repair work, accidents and weather issues, for instance – may alter track conditions significantly and cause outliers in the above-mentioned measurements without presupposing that the detector has failed, in fact there is a research field on this issue (among others Thomas and van Berkum (2009), Zhang et al. (2008), Srinivasan et al. (2007) Tang. and Gao (2005)).

The bibliography Lin et al. (2012) indicates several error detection techniques based solely on historical consistency. They do not take nearby detectors, that is, network consistency, into consideration. Other approach is to incorporate observations from nearby detectors Vanajakshi and Rillet (2004). This paper presents a method that is complementary to the existing ones, where basic consistency and network consistency are taken into consideration.

The method automatically detects a TCS that is failing, by only considering the data observed by the network detectors as input data. Once the detector that is failing has been detected, the procedure can be repeated to see if the remaining measurements are consistent and free of errors.

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 urban network; Section 4 discuss the effect of the model's variables on the results; and, finally, Section 5 presents the main conclusions of the paper.

4.3.2. Methodology

The method presented in this paper to detect and identify a malfunctioning detector is based on the resolution of a linear programming problem (LP).

In general terms, the \mathfrak{R}^n region that meets certain restrictions is known as the LP's feasible region. That is what will be built for the problem posed in this paper.

Feasible region

Let a series of measurements be taken $\{x_i^{obs}\}$ and that the tolerance indicated for each measurement is α_i . This tolerance is usually expressed as a percentage of the measured value, since it is reasonable to assume that any absolute errors incurred will be lower for small magnitudes than for larger ones, assuming the detectors function under the conditions specified by the manufacturer: $\forall i; x_i^* \in [a_i, b_i]$, where $a_i = x_i^{obs} - \alpha_i x_i^{obs}$ and $b_i = x_i^{obs} + \alpha_i x_i^{obs}$. In example 1, a 3% error was considered admissible for all the measurements, and therefore we would take $\forall i, \alpha_i = 3\%$, although in other cases a different error for each detector could be considered.

Given a set of observed values $\{x_i^{obs}\}$, $i \in I$, (where I is a set of indexes) each with a tolerance of α_i , we define the *admissible region* as the set $A \subset \mathfrak{R}^n$, such that $\forall \vec{x} = \{x_i\} \in A$ where the following conditions are satisfied:

1. $x_i^{obs} - \alpha_i x_i^{obs} \leq x_i \leq x_i^{obs} + \alpha_i x_i^{obs}$.
2. Vector \vec{x} verifies flow conservation laws.

Attention should be paid to the fact that the cardinal of the set of observed values and the number's n dimension do not necessarily coincide. Thus, to continue with the previous example, the set of observed values could be $x_1^{obs} = 800$, $x_2^{obs} = 1200$, $x_3^{obs} = 600$ y $x_4^{obs} = 700$, which would give the admissible region:

$$A = \{ \vec{x} \in \mathfrak{R}^5 / 776 \leq x_1 \leq 824; 1164 \leq x_2 \leq 1236; 582 \leq x_3 \leq 618; 679 \leq x_4 \leq 721; x_5 = x_1 + x_2 - x_3 - x_4 \}$$

Where the first 4 intervals are obtained by $x_i = x_i^{obs} \pm \alpha_i x_i^{obs} = x_i^{obs}(1 \pm \alpha_i)$ adding the flow conservation law: $x_1 + x_2 = x_3 + x_4 + x_5$.

Theorem 1 If all the detectors function properly, the feasible region is not empty ($A \neq \emptyset$).

Obviously, if all the detectors give admissible errors, then the true values vector belongs to the feasible region ($\vec{x}^* \in A$).

Therefore, the inference is:

Corolary 1 If $A = \emptyset$, one of the detectors is giving an inadmissible error.

Corolary 1 provides a method for detecting incorrect measurements by taking into consideration fundamental inconsistencies and network inconsistencies. Although the reciprocity theorem is not true, that is:

A detector may produce an inadmissible error, but the remaining detectors' margins permit admissible values and, therefore, $A \neq \emptyset$. In practice, this means that although a measurement's margin may be wider than required, it is not too inconsistent. So, if a detector is severely malfunctioning it will be impossible to generate consistent traffic counts.

We should also consider that if there are several vectors in A ($A \neq \emptyset$), some are more plausible than others, in so far as they are closer to the observed values. So, for a vector $\vec{x}^* \in A$ we can associate another vector $\vec{h} = \{h_i\}$ such that the verisimilitude of the i -th component is:

$$h_i^* = 1 - \frac{|x_i - x_i^{obs}|}{\alpha_i |x_i^{obs}|} \quad h_i = \max\{0, h_i^*\} \quad (16)$$

Figure 12 shows the verisimilitude of assigning a value x_i when x_i^{obs} with reliability α_i has been observed.

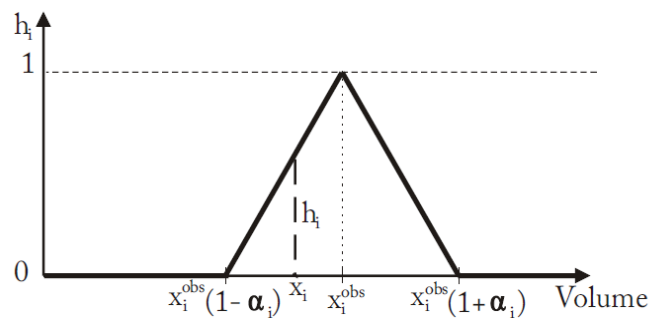


Figure 12. Verisimilitude function for a single observation

For the sake of simplicity, a triangular shape function has been chosen since the function shape is not an important issue, since the aim is to check if the adjusted value is in or out of the feasible region and simplicity of linear decay allows it to be solved by linear programming. However, other polygonal function could be used as it is stated in Kikuchi and Miljkovic (1999).

Assuming $x_i^{obs} > 0, \forall i \in I$ and making the relevant transformations in the above equation, finding out whether an admissible set of values exists becomes a problem of finding out whether a solution to the linear optimization problem exists:

Problem 1

$$\text{Maximize : } \sum_{i \in I} h_i$$

$$\text{Subject to } \begin{cases} 0 \leq h_i \leq 1, x_i \geq 0, \\ x_i + \alpha_i x_i h_i \leq x_i^{obs} (\alpha_i + 1) \\ -x_i + \alpha_i x_i h_i \leq x_i^{obs} (\alpha_i - 1) \\ M\bar{x} = 0 \end{cases} \quad (17)$$

where x_i , h_i and h are the variables that can be considered adjusted (consistent) values, variable verisimilitude and minimal verisimilitude, respectively, and where the flow conservation laws are represented by the homogeneous linear system $M\bar{x} = 0$. Thus, for case 1 with the single conservation law: $x_1 + x_2 - x_3 - x_4 - x_5 = 0$, the matrix $M = (1, 1, -1, -1, -1)$. In general, the matrix M will have as many rows as existing flow conservation equations. Very different target functions could have been selected for this task, but this will also serve the second aim of this paper: To determine which detector is producing erroneous values. The benefit of transforming the problem into a linear programming problem is being able to count on multiple and optimized routines for the solution. See Saameño et al. (2006). It is easy to amend the above method to consider different margins to the right and to the left of the observed values, i.e. $x_i^{obs} \in (x_i^{obs} - \alpha_i^L x_i^{obs}, x_i^{obs} - \alpha_i^R x_i^{obs})$.

Detection of inadmissible measurements

Let the problem of resolving linear programming 1 in the above section be posed and that there is no solution, since $A = \emptyset$. We would be in the case of Corolary 1, which indicates that one of the measurements is inadmissible. Unfeasible should not be confused with outliers, since the latter may be correct and due to traffic anomalies (an accident, repairs, etc.) but consistent with flow conservation laws.

To detect an incorrect measurement, we relax the manufacturer's α_i margins, multiplying them by a constant $K \gg 0$ so the new linear optimization problem will have a non-empty admissible region. That is:

Problem 2

$$\begin{aligned}
 & \text{Maximize : } \sum_{i \in I} h_i \\
 & \text{Subject to } \begin{cases} 0 \leq h_i \leq 1, x_i \geq 0, \\ x_i + K\alpha_i x_i h_i \leq x_i^{obs} (K\alpha_i + 1) \\ -x_i + K\alpha_i x_i h_i \leq x_i^{obs} (K\alpha_i - 1) \\ M\vec{x} = 0 \end{cases} \quad (18)
 \end{aligned}$$

It is known that one property of the 'maxsum' objective function is that it gives high values to most variables at the expense of giving low values to a few ones, Saameño et al. (2006). In this case, its effect is to assign values very close to the observed values (high verisimilitude) to the detriment of assigning very distant values to a few (low verisimilitude). The measurement that produces $h = \min\{h_i\}$ in problem 2 will be proposed as inadmissible.

We can always obtain a K that is large enough to make $A \neq \emptyset$; for its effect is to increase the variables' admissible margin. In an extreme case, any measurement x_i would fit into the $(x_i^{obs} \pm K\alpha_i x_i^{obs})$ interval. It could be assumed that selecting K would modify the solution obtained, but the following theorem shows that such is not the case:

Theorem 2 If the problem 2 is solved by using two different values for K ($K_1 \neq K_2$), performing both feasible solutions, then optimum solutions for K_1 and K_2 verify:

The optimum vector $\bar{x}^{(1)*}$ for K_1 is also optimum vector for K_2 : $\bar{x}^{(2)*} = \bar{x}^{(1)*}$

The index of observation with minimum value for h_i is the same for both constants: $\text{argmin}_i\{h_i^{(1)}\} = \text{argmin}_i\{h_i^{(2)}\}$

Proof is given in [Appendix](#) (shown in the published paper enclosed in the Annexe of this document)

Proposed algorithm.

From previous considerations, next algorithm is proposed.

Algorithm 1 (Erroneous sensor detector)

- 1) Read values for x_i^L , x_i^R y x_i^{obs} .
- 2) Represent the flow conservation laws by matrix M.
- 3) Repeat until all $h_i > 0$,
 - a) Represent all inequalities by matrix A and vector \vec{b} :

$$x_i + \alpha_i^R x_i h_i^* \leq x_i^{obs} (\alpha_i^R + 1)$$

$$-x_i + \alpha_i^L x_i h_i^* \leq x_i^{obs} (\alpha_i^L - 1)$$
 - b) Express restrictions $x_i \geq 0$
 - c) Solve LP with the target function *Maximize*: $\sum_i h_i^*$
 - d) If all $h_i^* \geq 0$, go to step 4, else:

d₁) Evaluate $h^* = \min_i h_i^*$, $K = \frac{1-h^*}{0.9}$

- d₂) Replace $\alpha_i^R \leftarrow K\alpha_i^R$ and $\alpha_i^L \leftarrow K\alpha_i^L$, into A and \vec{b} (step 2a)
- d₃) Solve LP with the target function *Maximize* : $\sum_i h_i^*$
- d₄) The index k that produces $h_k^* = \min_i h_i^*$ is obtained
- d₅) Observation x_k^{obs} is ellipsed and considered as erroneous
- d₆) Return to step 3). With initial values of α_i^R and α_i^L , but the ellipsed one.
- 4) Finish (Ellipsed observations are considered as inadmissible ones).

The algorithm is focused on detecting inadmissible observations from the network consistency viewpoint. However, it is easy to incorporate any available additional information. For instance, by changing the upper bound of any variable (adding the restriction $x_i \leq U_i$ in step 3b), or by changing the lower bound of any variable, which by default is 0 ($x_i \geq L_i$), etc. This fact allows making it suitable to perform fundamental consistency, generally expressed by bounds.

This method could be complementary to standard pre-process that analyzes historical consistency, Lin et al. (2010). That is, observed variables must be into a real interval, in other case the observation is considered an outlier. An outlier must be analyzed separately since it can be produced by anomalous traffic, but be correct.

Perhaps the algorithm 1 was only executed to verify that the detectors were working properly, but it is usually part of the study on a region's traffic. In such case, the next step would be to obtain the adjusted data, that is, the consistent data that most closely resembles the observed data. Any data deemed inadmissible during the pre-process will have been eliminated from the observed data using one of the procedures suggested by other authors as Kikuchi and Miljkovic (1999), Wall and Dailey (2003), Vanajakshi and Rilett (2004) and de Oña et al. (2011).

4.3.3. Application to an urban network

Road network data

The method is applied to the urban network shows in Figure 13.

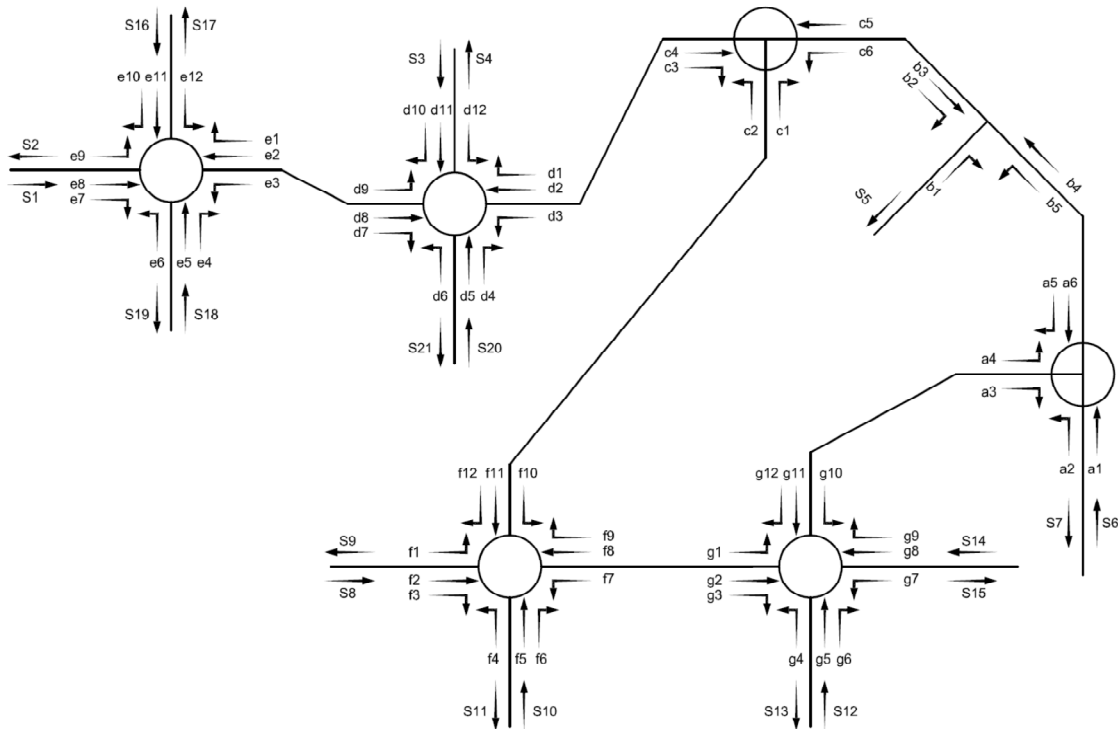


Figure 13. Example of an Urban Network

The network has seven intersections, of which four have twelve movements (intersections D, E, F and G), two have six movements (A and C), while the last one has five potential movements (intersection B). So, in total, there are 86 unknown variables. Since it is impossible to guarantee that a set of true data will always be available, the initial set of data will be a set of consistent data that is very close to the observed data.

Consider the situation shown in Figure 13, in which consistent true data are available (Theoretical Values – TV), where the data that comply with flow conservation in the traffic network concerned is deemed to be consistent. In other words, the sum of incoming vehicles is equal to the sum of outgoing vehicles at any network intersection.

This consistent data-base is used to randomly deform values with a tolerance of $\pm 3\%$ (by an uniform distribution), which is the tolerance shown by the count stations most commonly used in urban networks, Cegasa, (2010). This is not deterministic, however, because if the detector was of another type or had a different tolerance, a value other than $\pm 3\%$ could be considered. The model allows a different α for each observed datum to be defined (several types of detectors with different tolerances). As shown in section 2, it even permits the definition of asymmetric feasible regions.

Having obtained a randomly distorted data base within the above-mentioned tolerance, it could then be considered as the data that would be obtained in an ideal counting campaign in which all 86 potential movements would be measured. Therefore, it could be taken as the series of observed data in an urban network (Observed Values – OV). In this case, the values would not be consistent, according to the above definition (the sum of incoming vehicles would not be equal to the sum of outgoing vehicles).

The fact that a base of consistent true data is used and subsequently randomly distorted permits a comparison between the results obtained and real life, and verification of the goodness of the method proposed.

Results

OV obtained randomly from TV with a tolerance of $\pm 3\%$ is used to verify the goodness of the model. Next, a datum is randomly selected and distorted to simulate a detector error that exceeds the error specified by the manufacturer or, in other words, a deviation from the detector's allowed tolerance. Specifically, deviations of 75%, 50%, 25% and 10% from OV are simulated.

This deformation gives an initial data base for each example generated (each of which contains an erroneous datum). For each one of the 4 deviations, 500 examples are randomly generated from OV. In all, 2,000 examples are executed. Table 6 shows the results for the random examples.

1	2	3	4	5	6	7	8
Percentage simulated error	Error is detected		Error is pointed out and gets by h_{min}		Error is pointed out and gets by 2 nd h_{min}		Error is pointed out in total
	Number of times (A)	Proportion A/500	Number of times (B)	Proportion B/500	Number of times (C)	Proportion (C)/500	Success proportion (B+C)/500
75	491	0.982	443	0.886	23	0.046	0.932
50	486	0.972	386	0.772	37	0.074	0.846
25	438	0.876	269	0.538	27	0.054	0.592
10	266	0.532	137	0.274	24	0.048	0.322

Table 6. Results with a simulated error of 75, 50, 25 and 10%

Column 1 in Table 6 shows the simulated error in a randomly selected measurement apparatus. Columns 2-3 show the number of times an error is detected in all the random samples. That is, the number of times $A=\phi$ is obtained applying theorem 1. Column 2 points out the number of times $A=\phi$ is obtained for the random examples, which is when the adjusted value lies outside of the detector's allowed tolerance, and outside of the set boundaries of the feasible region. This indicates that a detector is giving a value that is higher than the allowed deviation, which in turn means that a detector is failing. Column 3 shows the same thing in relative terms.

By increasing α from 0.03, (see d1) in the algorithm 1) in an iterative process, the feasible region is extended in order to allow $A \neq 0$ to be found for every i . This value was selected because it was considered a sufficiently large amount for adjusted values to be found in the feasible region. If any h_i equal to 0 was found when linear programming was executed, the constant would be increased even further, since it would not affect the results, as justified by Theorem 2.

Table 6, column 4 shows the number of times the index i that produces $h = \min h_i$ coincides with the failing TCS. Columns 6-7 show the number of times (and proportion, respectively) in which the failing TCS is the one that shows the second lowest value. So, when a TCS perform a 75% of error, it coincides with the error obtained by the second minor value of h_i in 5% of cases.

Column 8 shows the proportion of times that the model is able to detect the failing detector (i.e. adding the number of times it detects the detector that fails, whether it is the h_i minimum or the value immediately above it). This result points out the proportion of times at which it indicates a detector that is failing, out of all the random examples. This is the model's proportion of success, and it is calculated by adding column 4 and column 6, and dividing by the total number of examples simulated. For an error of 75%, the success rate is 93%. For the remaining cases, the model finds that there is a malfunctioning detector, but it does not point it out in the first or second places.

Table 6 shows that the model's success increases in the same measure as the device's error increases and worsens as the error diminishes, and the closer it is to the measurement device's tolerance range.

If the ratio (r) is expressed as the proportion of times that an error is detected compared to the number of examples executed (Table 6, column 2), the ratio of cases in which a failing detector is detected for each simulated error can be compared.

In other words, if N random examples have been executed (in this case, $N = 500$) and A times errors have been detected (Table 6, column 2), the estimated ratio obtained experimentally is $r = \frac{A}{N}$ (Table 6, column 3). In this manner, for an error of 75%, the error is detected in 98% of cases; for an error of 50%, in 97% of cases; for an error of 25%, in 88% of cases; and finally, for a simulated error of 10%, an error is detected in 53% of cases.

4.3.4. Sensibility analysis to different variables

First, the effect of the situation of the failing traffic counts will be analyzed. Second, what happens when certain points of the network have not been counted? Finally, the sensitivity to the number of not counted data in the network will be analyzed (with approximately 50% more and 50% less points not counted).

Percentage simulated error	Error is detected		Error is pointed out and gets by h_{min}		Error is pointed out and gets by $2^{nd} h_{min}$	
	center	edge	center	edge	center	edge
75	492	500	426	500	24	0
50	474	500	340	486	49	6
25	419	500	227	428	32	25
10	205	466	90	249	11	45

Table 7. Results for center and edge detectors with a simulated error of 75, 50, 25 and 10%

Effect of the situation of the failing detector

How does the sensitivity depend on which detector is malfunctioning? In order to analyze the method's sensitivity to the detector position, a selective choosing of the malfunctioning detector has been made. At first stage, for each scenario, the model was forced to choose an edge detector, (S_1, S_2, \dots, S_{21} , or b_1 in Figure 13), and at second stage the central ones (the remaining detectors) have been chosen to be failing. Table 7 shows the results.

The method detects an error on the edge of the network better than when the detector is situated in the center. This is logical due to the following reason: when an edge detector is getting an inadmissible error, while the rest adjacent measurements are corrects, must significantly modify its value in order to reach network consistency. That is because a small amount of adjacent detectors exists which can be modified within the margin established by the feasible region. On the other hand, a major modification of these adjacent detectors makes the constraints able to be affected; therefore the $\sum_i h_i$ is reduced. The target function forces to modify the one that is giving an erroneous measurement.

While, for center detectors, the measurements are linked to more variables that can be modified within the feasible region. So, for an inadmissible small error (around 10%) is easier to count on the adjacent values margin and move all of them, in order to get all measures within its feasible region, than a big change in the malfunctioning detector.

Effect of points that are not counted

In this subsection the effect of movements that have not been counted is analysed.

Presumably, the network in Figure 13 shows seven movements that have not been counted (movements c_2 , c_3 , c_4 , d_9 , d_{10} , d_{11} and d_{12}). This implies around 8% of all the movements in the network. This percentage is considered normal in counting campaigns in a traffic network (Zhong et al, 2004). A case consisting of 500 random examples is simulated below, in which the number of not measured movements is increased 50% (10 not measured movements), followed by a case in which the number of not measured points is diminished in 50% (4 not measured movements) .

Table 8 and Figure 14 show a comparison between the results obtained in the study with 4 hypotheses (all measured data, 4, 7, and 10 not measured data). In Figure 14 the x-axis represents the simulated distortions for the measurement device and the y-axis represents the proportion of times the error is detected.

Percentage simulated error	Error is detected		Error is pointed out and gets by h_{\min}		Error is pointed out and gets by $2^{\text{nd}} h_{\min}$		Error is pointed out
	Number of times (A)	Proportion A/500	Number of times (B)	Proportion B/500	Number of times (C)	Proportion C/500	Success proportion (B+C)/500
4 not measured movements (50% less)							
75	466	0.932	397	0.794	27	0.054	0.848
50	447	0.894	353	0.706	36	0.072	0.778
25	391	0.782	234	0.468	38	0.076	0.544
10	242	0.484	95	0.190	29	0.058	0.248
7 not measured movements							
75	449	0.898	374	0.748	30	0.060	0.808
50	425	0.850	315	0.630	42	0.084	0.714
25	351	0.702	201	0.402	28	0.056	0.458
10	226	0.452	100	0.200	30	0.060	0.260
10 not measured movements (50% more)							
75	418	0.836	360	0.720	22	0.044	0.764
50	372	0.744	282	0.564	35	0.070	0.634
25	338	0.676	211	0.422	29	0.058	0.480
10	203	0.406	87	0.174	28	0.056	0.230

Table 8. Results with a simulated error of 75, 50, 25 and 10% with four, seven and ten not measured movements.

Taking column 3 (A/500) in Tables 6 and 8 into consideration, a comparison can be made about the number of times an error is detected in each case. Table 8 shows that the ratio of errors detected for the simulated scenarios gradually diminish when there is less measured data available (i.e. less information).

From Figure 14, Tables 6 and 8, it is possible to analyze the model's sensitivity to the number of not measured movements in a case where all the data from all the TCS (i.e. all measured data) is available.

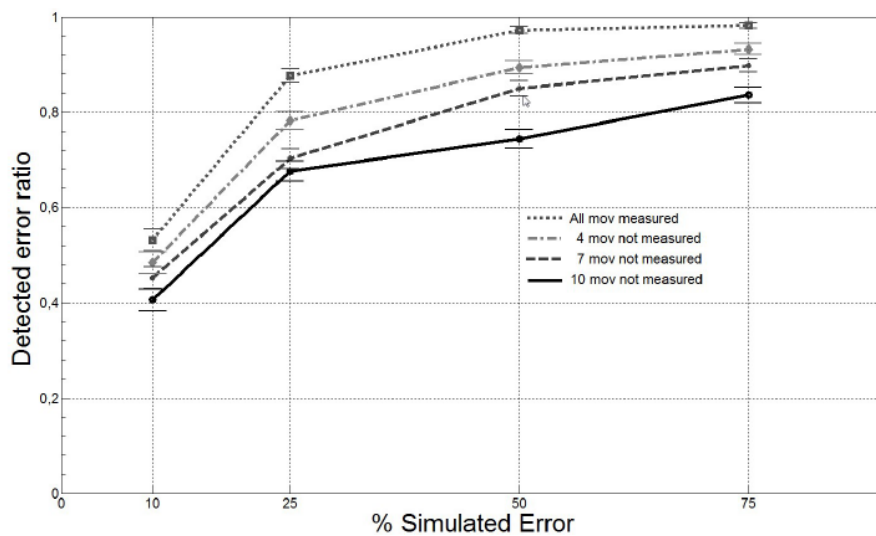


Figure 14. Sensitivity analysis of success versus increase in the number of data not measured.

The x-axis represents the simulated percentage of the device error (10, 25, 50 and 75%) and the y-axis data shows the percentage of success for every case, in comparison with the one in which all the data are measured. In the event that a 75% error occurs in a detector, for instance, the chart will show that the model presented in this paper is 93% successful if 4 network data are not measured, 90% if 7 network data are not measured, and 84% if 10 data are not measured.

Thus, the conclusion would be that the model gives good success results even when the number of not measured data increases, although, obviously, when more data is available, the more it improves.

Combined effect of the size of the error and the number of points that are not counted

Figure 14 shows the ascending trend of the ratio when a detector's error increases in all the hypotheses. The trend is even more pronounced when it moves from an error close to the detector's tolerance range (such as 10%) to around 25%, after which the detector's behaviour is asymptotic, reaching an error ratio within the range 0.9-1 for the biggest device error simulated. In other words, when the error exceeds the threshold at around 25%, it can be asserted that the model succeeds in around 90% of the cases.

In figure 14 the $1-\sigma$ errors bars have been included in order to show conclusions do not owe to random.

Percentage of simulated error	All movements are measured		4 not measured movements		7 not measured movements		10 not measured movements	
	Proportion	σ	Proportion	σ	Proportion	σ	Proportion	σ
75	0.982	0.006	0.932	0.011	0.898	0.014	0.836	0.017
50	0.972	0.007	0.894	0.014	0.850	0.016	0.744	0.020
25	0.876	0.015	0.782	0.018	0.702	0.020	0.676	0.021
10	0.532	0.022	0.484	0.022	0.452	0.022	0.406	0.022

Table 9. Ratios calculated for every scenario providing the standard deviation

Table 9 showed the ratios (or proportion of success, p_i) at which error is detected in every scenario. To demonstrate that the model's proportion of success increases when more data are measured ($p_{i+1} < p_i$) and that the observed results are not due to chance, a hypotheses of proportional difference was tested at a significance level of 5%, taking $N_{i+1} = N_i = 500$. See Anderson and Sclove, (1986) and Mendenhall and Sincich (1988).

Three statistical tests were conducted to compare the three hypotheses in groups of two. That is, firstly hypothesis of all movement measured was tested versus 4 not measured movements, the case of 4 not measured movements versus 7 not measured data, and lastly, 7 not measured data were tested versus 10 not measured data. The $Z_{exp} = \frac{p_i - p_{i+1}}{\sigma}$ is calculated and compared with the $Z_{theoretical} = 1.645$, it determines the significant region ($Z_{exp} > 1.645$). The results are given in Table 10.

It is found that $p_{i+1} < p_i$ in all cases and statistically significant results are obtained for the cases of 75, 50 and 25% error in the first and second tests, and in the third one the results are significant after the 50% error.

Therefore, it can be asserted that the success proportion improves with the number of counted data and this fact is not due to chance.

% error	p_i	p_{i+1}	Z_{exp}
All movements measured vs 4 not measured movements			
75	0.9820	0.9320	3.92711065
50	0.9720	0.8940	4.99384767
25	0.8760	0.7820	3.97862436
10	0.5320	0.4840	1.51983992
4 vs 7 not measured movements			
75	0.9320	0.8980	1.93124468
50	0.8940	0.8500	2.0869071
25	0.7820	0.7020	2.90315821
10	0.4840	0.4520	1.01452938
7 vs 10 not measured movements			
75	0.8980	0.8360	2.89896925
50	0.8500	0.7440	4.20341134
25	0.7020	0.6760	0.8884334
10	0.4520	0.4060	1.47112841

Table 10. Test of hypotheses. Significant cases in bold.

4.3.5. Summary and Conclusions

This paper presents a method for detecting inadmissible errors in TCS and identifying which device is more likely to be failing. The method is based on a double linear optimization process that can easily be solved with existing software on the market, and which we consider highly useful for practitioners.

If the method detects the existence of an inadmissible error in the TCS' measurements when the first linear optimization is used, a second optimization can be used so the method can obtain the detector that is most likely to be failing (the one that obtains the $\min_i h_i$). This facilitates to replace or fix them for obtaining adjusted data.

Four different cases of potential errors were simulated in order to identify the effects on the method (deviations of 10%, 25%, 50% and 75%). The results

show that the method works better with bigger errors (75%), which are more frequent when dealing with malfunctioning detectors, than with small errors (10%), close to the TCS's tolerance (3%). For deviations of around 25% of their theoretical value, the method is 88% efficient for detecting that there is an error in the measures. The efficiency in identifying a failing detector can be considered good (over 90%) when the error is over 75% of the deviation, and diminishes as the errors become smaller.

The same tolerance was considered for all the TCS (3%), but the model is versatile and allows assigning a different tolerance to each detector according to its type and level of precision.

Finally, a statistical test has been conducted to demonstrate that the increase in the number of times an error is detected when more movement counts were obtained as opposed to a gradually decreasing number of times is not due to chance. This serves to assert that the results are significant and the size of the sample selected is sufficient to corroborate the conclusions arrived at in this paper.

Usually studies perform automated data checking by comparing measured data to historical data for consistency Lin et al. (2012). Sometimes, however, there are no historical data and only the observed database is available. This is when the method proposed in this paper becomes a good tool for detecting errors, since the only incoming data required are the observed data, with no need for preprocessing. Actually, both approaches could be considered as complementary: it is possible to use fundamental and network consistency for detecting inadmissible errors and, historical consistency as alarm signal.

4.3.6. Acknowledgment

The authors appreciate the reviewers' comments and effort in order to improve the paper.

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

De Oña, J., Gomez, P. and Merida-Casermeiro, E. (2013). Adjustment boarding and alighting passengers on a bus transit line using qualitative information. *Applied Mathematical Modelling* <http://dx.doi.org/10.1016/j.apm.2013.07.041>.

For 2012, the journal APPLIED MATHEMATICAL MODELLING has an Impact Factor of 1.706. This table shows the ranking of this journal in its subject categories based on Impact Factor.

Category Name	Total Journals in Category	Journal Rank in Category	Quartile in Category
Engineering, Multidisciplinary	90	11	Q1
Mathematics, interdisciplinary Applications	92	18	Q1
Mechanics	134	32	Q1

The preliminary results of this work were presented at the X Transport Engineering Conference (CIT2012) held in Granada, June 20-22th 2012.

Abstract

Obtaining data to use in an urban public transport operation planning and analysis is problematic, specifically in urban bus transit lines. 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 O/D of passengers allowing the O/D matrix calibration by using the loads among 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 transit line in Malaga (Spain) and to a set of 50 different 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%.

Keywords: fuzzy optimization; transport planning; public transport

4.4.1. Introduction

When planning public transport, it is crucial to know the real O/D of passengers. Surveys about the starting point and destination of travelers are mandatory to obtain the real O/D matrix at every transport system. Once the former is configured, it has to be calibrated with collected or measured data. For that aim, the loads among transit line stops are used, and to get them, in and out movements of passengers at each stop along a transit line are required. On the other side, loads become crucial in the 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. (2011), and thanks to the increased sophistication of mobile communication technologies (Blythe, 2004). Smart cards improve the quality of data (Dempsey, 2008) and the ticket validation systems provide information on the number of boardings. Therefore, the information is quite accurate and the only errors are due to potential device failures.

However, the systems cannot be used to obtain data on the number of alightings, so passenger detection systems and surveyors on board or at the stops are needed for that purpose. Several surveyors may be needed if there are several exits 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 (Qing et al., 2009) (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 ticket before boarding and in most cases in underground network 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 bus urban or metropolitan transit lines worldwide.

4.4.2. Background

Several methods have been developed to adjust data on a transit line when both boarding and alighting data are available (Kikuchi et al., 2000). 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 (Kikuchi and Miljkovic, 1999). 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 (Chen, 2002).

• 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 don't 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.

4.4.3. Theoretical Approach

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:

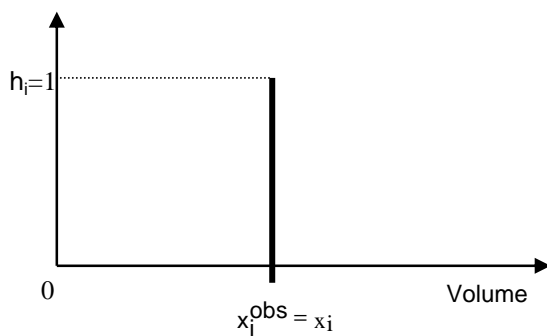
- the total number of boarding passengers should be equal to the total number of alighting passengers, and

- 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})

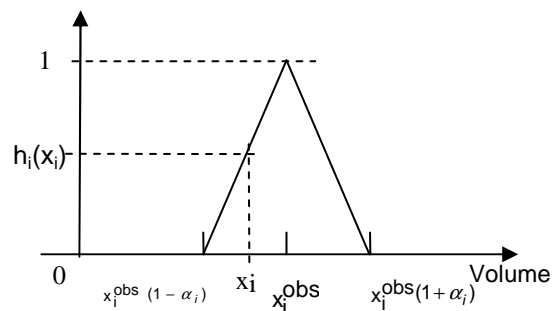
The initial variables and data for solving the problem are the values for passenger boardings and alightings, vehicle loads between stops and L_{\max} .

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).

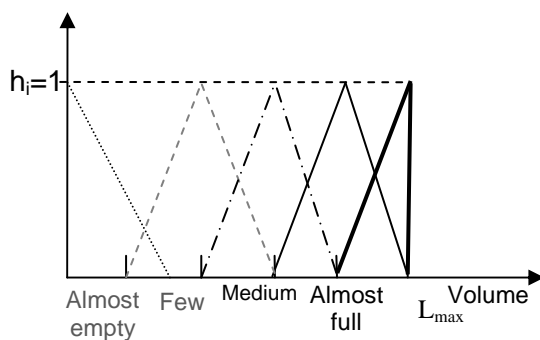
e) Fix Number



f) Crisp Number



g) Fuzzy Information



h) Missing Value

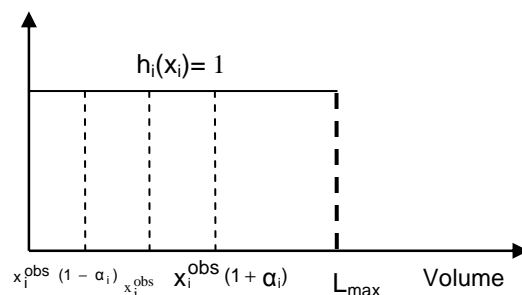


Figure 15. Membership functions for (a) fix number, (b) crisp number, (c) fuzzy information and (d) missing value

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 and 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.).

The proposed method

The first step to solve this problem is to use membership functions to represent the above concepts. Figure 15 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 (1990), Yager and Filev (1994), Zimmermann (1996), and Klir and Wierman (1999).

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 (Lotfi et al., 2009). 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. 19 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.

$$h_i(x_i) = \max\left\{0, 1 - \frac{|x_i - x_i^{obs}|}{\alpha_i x_i}\right\} \quad (19)$$

x_i is the adjusted value for the i -th variable. That is: $\forall i; x_i \in [x_i^{obs} - \alpha_i x_i^{obs}, x_i^{obs} + \alpha_i x_i^{obs}]$. 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 Figure 15 are specific cases of case (b). A fixed number $\alpha_i=0$ forces its value to be kept after the adjustment, i.e. $x_i^{obs} = x_i$. 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 $\{x_i^{obs}\}_{i \in I_b \cup I_a \cup I_L = I}$, (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 $A \subset \mathbb{R}^n$, such that $\forall \bar{x} = \{x_i\} \in A$ where the following conditions are satisfied:

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 \bar{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. (2011):

Step 1: The problem is solved using MaxMin Method (MM method), (Eq. 20), and we obtain a value of $h = \min(h_i)$.

$$\text{Max}(h) \text{ where } h \text{ is } \min(h_i) \quad (20)$$

Subject to

Constraints related to the membership functions:

$$h_i(x_i) \geq h \quad \text{for } i=1, 2, \dots, 3N \quad (21)$$

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:

$$\sum_{i \in I_b} x_i = \sum_{i \in I_a} x_i; \quad \text{for } i=1, 2, \dots, 3N, \quad (22)$$

where N is the number of transit stops

Constraints related to vehicle conditions:

$$l_k \geq 0 \text{ and } l_k \leq L_{\max} \quad (23)$$

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. 24):

$$\text{Max}(g) \text{ where } g \text{ is } \sum(h_i) \quad (24)$$

Subject to the same constraints related to the conservation of flow at the transit line (Eqs. 22 and 23), and to the following constraints related to the membership functions:

$$h_i(x_i) \geq h^* \quad \text{for } i=1, 2, \dots, 3N \quad (25)$$

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 (Linprog, 2011; Lotfi et al., 2009).

4.4.4. 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.

Example 1: transit line in Malaga

Figure 16 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 11 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. 22).

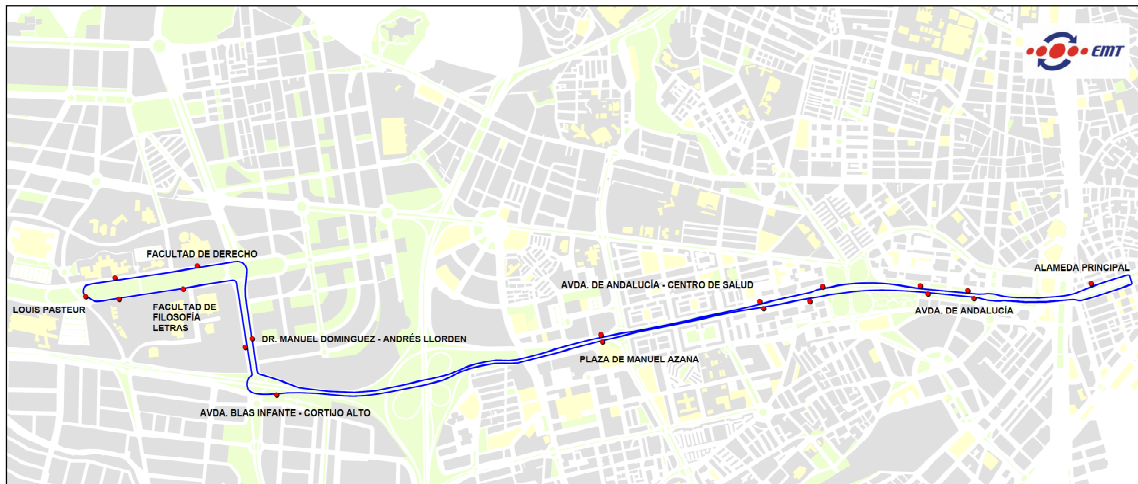


Figure 16. Example of a transit line in Malaga

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

Table 11. Alightings and boardings true values for a transit line in Malaga

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 1,000 times. Therefore, from the true consistent data x_i^{true} (see Table 11), 1,000 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.

Examples to 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:

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. 22) and related to vehicle conditions (Eq. 23) are imposed. This database is used to verify the goodness of fit of the method (see following Sections).

In every fictitious transit line, the consistent data generated is randomly deformed in the same way and with the same tolerance as it was for the transit line in Malaga (see Section above): $\pm 25\%$ 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 Eq. 22 and 23, and are considered as the data that would be obtained during a conventional data collection, and they are the input for the model.

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, to generalize and validate the proposed method we will apply it to 5,000 different transit lines with a number of stops between 10 and 75.

Scenarios

As pointed in the above Section, 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:

No further qualitative information is available on the remaining alightings and loads: missing alighting (MA) and missing loads (ML)

Qualitative information is available on the alightings (FA) where no quantitative information exist

Qualitative information is available on the vehicle loads (FL) between successive stops

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 12). To analyse the 5,000 transit lines for generalization and validation of the method, 12 scenarios are considered (bold scenarios in Table 12).

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 12. Scenarios definition

In Table 12, 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.

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 12).

In both cases under study (one in the case of the transit line in Malaga; 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. 26 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.

$$\varepsilon = \frac{\sum_{i=0}^n |x_i - x_i^{true}|}{n} \quad (26)$$

The average error, the standard error deviation, the minimum and maximum error can be obtained from ε . Table 13 shows the average errors obtained from ε committed in the 1,000 defined cases in Example 1, under the 40 different scenarios. Furthermore, this Table 13 also shows the average errors when LSM is used under the 10 aforementioned scenarios.

Table 13 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 5,000 cases under study for the 12 different scenarios. Table 13 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.

4.4.5. 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. Figure 17 shows the membership functions of the load and of the alightings in a bus carrying 100 passengers.

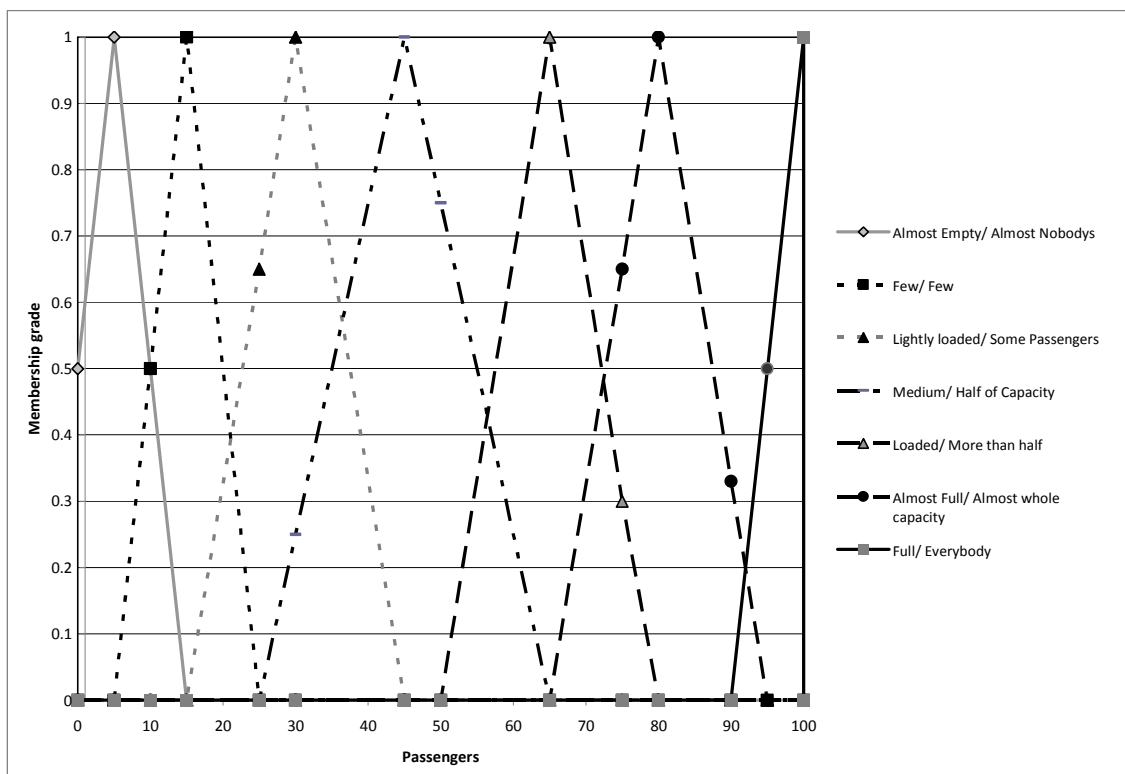


Figure 17. Membership functions of loads and alightings in a transit line

Table 13 shows the results for the 40 scenarios in the transit line in Malaga. The values in Table 13 represent the average error ($n=1,000$) for each scenario. Figure 18 shows the results in Table 13 graphically.

LSM		Cases ML/MA		Cases ML/FA		Cases FL/MA		Cases FL/FA	
CASE	ϵ	CASE	ϵ	CASE	ϵ	CASE	ϵ	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

Table 13. Results for the 40 different scenarios of 1,000 examples

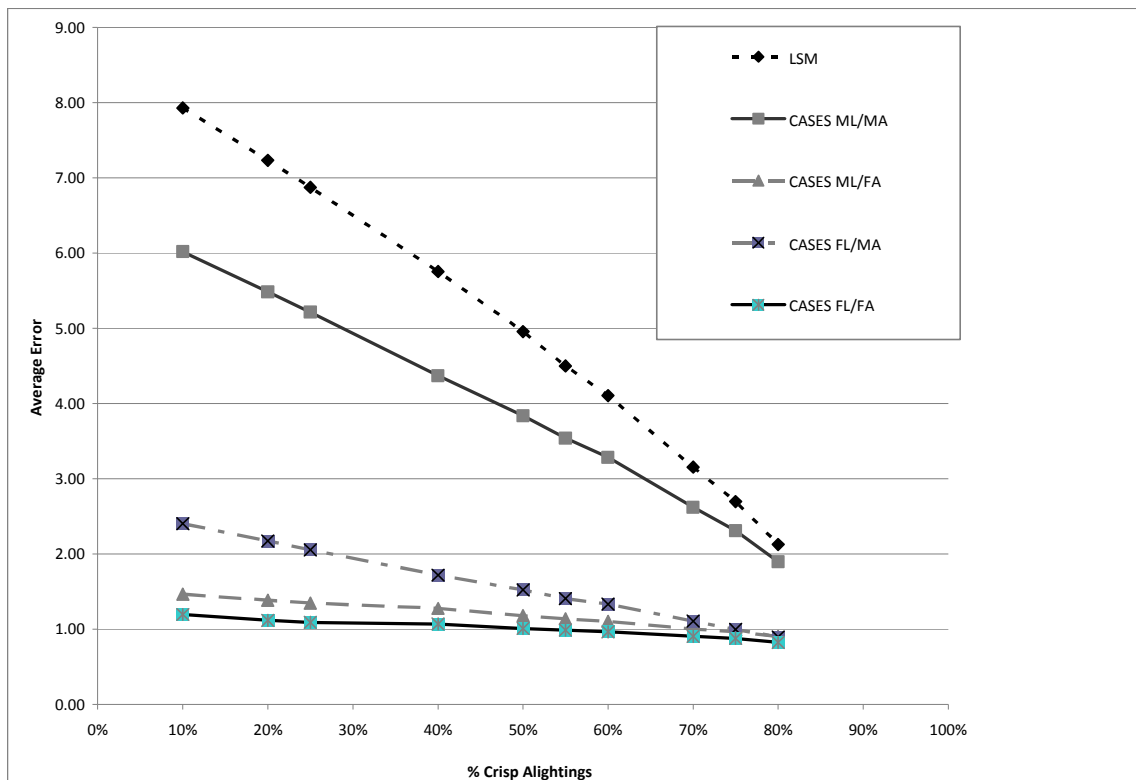


Figure 18. Error evolution

Table 13 and Figure 18 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 Figure 18 is much greater when only 10% of crisp alightings are available than when 80% are available.

Results in Table 13 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 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 14 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=5,000) under the 12 scenarios considered (bold scenarios in Table 12). In order to compare the results, the 3 scenarios that do not consider qualitative information have been adjusted by using the LSM. 15 cases are compared in total (3 scenarios with LSM and 12 scenarios with the proposed method).

	%crisp alightments	%fuzzy alightments	fuzzy loads	No. Cases	Average Error	Standard Dev	Min	Max
Total				75,000	4.86	2.32	1.00	15.33
Least Squared Method (LSM)								
ML/90MA	10%	0%	N	5,000	9.95 a	1.47	5.76	15.33
ML/60MA	40%	0%	N	5,000	7.98 b	1.35	4.16	14.33
ML/30MA	70%	0%	N	5,000	5.61 c	1.10	2.22	11.89
Proposed Method								
ML/90MA	10%	0%	N	5,000	7.59 d	1.01	4.13	12.70
ML/60MA	40%	0%	N	5,000	5.90 e	0.84	2.80	10.78
ML/30MA	70%	0%	N	5,000	4.27 f	0.68	1.67	8.00
ML/90FA	10%	90%	N	5,000	3.45 g	0.53	1.83	6.60
ML/60FA	40%	60%	N	5,000	3.12 h	0.53	1.42	6.10
ML/30FA	70%	30%	N	5,000	2.75 i	0.53	1.00	6.10
FL/90MA	10%	0%	Y	5,000	5.70 j	1.11	2.24	10.78
FL/60MA	40%	0%	Y	5,000	4.43 k	0.87	1.52	8.24
FL/30MA	70%	0%	Y	5,000	3.37 l	0.65	1.25	6.24
FL/90MA	10%	90%	Y	5,000	3.06 m	0.49	1.55	5.46
FL/60MA	40%	60%	Y	5,000	2.90 n	0.49	1.45	5.10
FL/30MA	70%	30%	Y	5,000	2.74 i	0.49	1	5.10

Table 14. Results (average error, standard deviation, min, max, and one-factor ANOVA) for 3 scenarios with LSM and 12 scenarios with the proposed method (n=5,000)

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.

	No.Cases	Average Error
Total	60,000	4.11
Fuzzy Alightments		
NO	30,000	5.21 a
YES	30,000	3.00 b
Fuzzy Loads		
NO	30,000	4.51 a
YES	30,000	3.70 b
Crispy Alightments		
10%	20,000	4.95 a
40%	20,000	4.09 b
70%	20,000	3.28 c
Fuzzy Alightments / Fuzzy Loads		
NO/YES	15,000	4.50
NO/NO	15,000	5.92
YES/YES	15,000	2.90
YES/NO	15,000	3.10
Fuzzy Alightments / Crispy Loads		
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
Fuzzy Loads / Crispy Alightments		
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 15. Results of two-factor ANOVA for the proposed method

Table 15 shows the two-factor ANOVA results. For this analysis LSM results are not considered. Table 15 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 15 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 15 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.

4.4.6. 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 slightly 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 present 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 and which are the zones where they go more loaded.

5 Chapter 5: CONCLUSIONS AND FUTURE RESEARCH

CHAPTER 5 Conclusions and future Research

5.1. General Conclusions

The general objective of this thesis is to propose a new method for adjusting field data in order to achieve consistency with regard to uncertainty, ambiguity and subjectivity. The consistency of observed traffic data is a concern because in nearly all cases traffic data contain errors. Processing observed data for consistency is crucial in any analysis where data interrelationships are important.

A new fuzzy optimization model has been developed so that subjective information can be incorporated. Its theoretical formulation and computational procedure are shown in the first paper published and are also presented in Chapter 3. The proposed approach is robust enough to deal with other common data discrepancies in transport situations. As has been shown, the model preserves the integrity of observed data as far as possible, and allows the analyst to distinguish between reliable and less reliable data.

Another contribution is in the field of data imputation and Base Data Integrity. The principle of Base Data Integrity is an important topic discussed by the American Society for Testing and Materials (ASTM, 1991) and the American

Association of State Highway and Transportation Officials (AASHTO, 1992). The principle states that traffic measurements must be retained without modification and adjustment. Missing values should not be imputed in the base data. However, this does not prohibit data imputation at the analysis stage. In some cases, traffic counts with missing values could be the only data available for certain purposes and data imputation is necessary for further analysis. In others, the devices used to collect the data may be malfunctioning and a method is developed to resolve these situations; this brings us to the second part of the thesis.

The method detects inadmissible TCS errors and identifies which device is most likely to be faulty. The method is based on a double linear optimization process that can easily be performed with existing software on the market, and which we consider highly useful for practitioners.

If the method detects an inadmissible error in the TCS measurements when the first linear optimization is performed, a second optimization can be carried out in order to identify the detector that is most likely to be malfunctioning. It can then be replaced or fixed so as to produce adjusted data.

The usual solution in such cases is to perform automated data checking by comparing measured data to historical data for consistency. A problem arises when there are no historical data and only the observed database is available. This is when the proposed method developed in this thesis becomes a good tool for detecting errors, since the only incoming data required are the observed data, with no need for pre-processing. Actually, both approaches could be regarded as complementary: it is possible to use fundamental and network consistency for detecting inadmissible errors, and historical consistency as an alarm signal.

To conclude the research work, the fuzzy optimization model was used to solve the problem of adjusting passenger boarding and alighting figures at each transit stop. This information is basic and must be used in the analysis of urban bus operations to calculate loads and calibrate O/D matrices obtained from surveys. The problem solved here is that observed counts of boardings and

alightings often do not match and also that alighting data are rarely available for urban bus systems today. The literature gives several different methods that are used to adjust boardings and alightings so that the basic principles of flow conservation are met. The methods are characterized by the need for numerical information (crisp values) in order to make the adjustment and the fact that the information must be obtained by automated or manual counts. Therefore, the effort required tends to be considerable.

In this part of the thesis a method is proposed that allows boardings and alightings on a transit line to be adjusted on the basis of qualitative information, consisting of the driver's, observer's or analyst's perception of vehicle loads between stops and the number of passengers who alight at each stop. This information can be obtained cheaply by public transport companies since, by having a quick look at the vehicle, the driver can choose one of the options that the analyst has defined beforehand (e.g. empty, almost empty, etc.).

The proposed method therefore works in those cases where other methods provide no solution because there are no available means to count the number of passengers who alight at each stop.

The method allows data adjustments to be made in cases where counts are possible but some data are missing, thereby avoiding the need to measure the data on the public transport line all over again.

From an operational point of view, this work also presents a new way to obtain information about loads between stops in order to regulate the service and improve it by adapting it to demand during peak times, thus making it easier to know when additional vehicles are required.

The main research objectives described in section 1 have been achieved:

- Since numerical information is highly costly and sometimes difficult to obtain, it is possible to make use of subjective information to deal with the uncertainties of field data, by using fuzzy logic optimization.
- A new fuzzy method has been developed to deal with field data uncertainty and to provide a database of adjusted values that are

consistent, by taking subjective information provided by the analyst into account.

- The proposed model produces a better adjustment of field data than the classical models. In fact, it has been proved that the classical methods are not able to deal with subjective information and so cannot reach a solution. Furthermore, the proposed model achieves better results than existing models that use fuzzy logic optimization, as shown in Chapters 3 and 4.
- Where malfunctioning TCS have to be detected, the proposed model is able to detect the faulty TCS with no additional information, while the existing methods require historical data to identify the error.
- The model is applied to solve operational problems in public transport planning, where counting alighting passengers is expensive and the collected information is very limited because such fieldwork is only conducted for a short period of time and only on a sample of the transit lines in the city studied. Using subjective information on the loads and alightings between consecutive stops to adjust the boarding and alighting figures solves this problem.

5.2. Future Research

Throughout the period in which this work was being researched and drafted, new lines of research continually arose but it was not possible to include them in this thesis. This section describes various lines of research that are currently under way, as well as others that are scheduled to begin.

Since information on traffic/passenger flows between specific origins and destinations in a transport network is the main kind of information required by planners and engineers for effective traffic management and control, O/D matrices are of vital importance for transport system planning and design, as well as for analysis, modelling and simulation. This is because of the information they contain about the spatial and temporal distribution of

movements between different traffic zones in an urban area (i.e. each cell represents the number of trips between an origin and a destination). O/D matrices are used to represent the current demand for transport systems, or, in conjunction with anticipated economic and population growth, land-use changes and planning policies, to identify and forecast future demand and other alternative scenarios. Future lines of research are therefore pursuing this aim, which is to forecast the O/D matrix for any transport system, whether a road network or a public transport system.

In the case of road networks, research is currently under way into the estimation of O/D matrices using loop detector data in combination with floating car data (FCD). We may introduce a method for doing this based on a bi-level optimization (BO) model using fuzzy logic theory. This data combination is rather promising and could be highly valuable for identifying not only demand patterns but also other more operational aspects of traffic. Furthermore, we may also try to establish the evaluation of trade-offs between FCD penetration rate and loop detector coverage for different accuracy levels in the estimation of O/D matrices. On the same subject, other research lines could use alternative input data, making use of newly emerging data collection technologies.

With regard to passenger flows in a public transport system, there may be another field of research to which the method may be applied, which involves creating a seed O/D matrix with the same input data described in the third paper, using redundant fuzzy information on loads and alightings, and in a second step updating it with a potentially available historical database of boardings and alightings. Additionally, the methodology proposed in this thesis could complement emerging technologies based on image recognition using street camera imaging; this could be useful to identify parking space availability, among other uses. Cameras are being installed in suburban railway stations to analyse transfers between lines and on urban buses to count alighting passengers. In any of these cases, the method created could be adapted for use with these new inputs to achieve the desired objective. The latest stage in the process, and one that we are planning to achieve soon, is to improve the method with artificial neural network techniques.

In conclusion, the methodology is versatile enough to be adapted and applied in any situation where uncertainty and ambiguity underlie the input data and represents a step forward in the planning and processing of input data for solving major transport problems.

REFERENCES

LIST OF REFERENCES

AASHTO Guidelines for Traffic Data Programs, 1992 American Association of State Highway and Transportation Officials.

Akiyama, T., Nakamura, K. and Sasaki, T., 1993. Traffic diversion model on urban expressway by fuzzy reasoning. Selected Proceedings of the Sixth World Conference of Transport Research, Lyon 92, pp. 1011-1022.

Akiyama, T. and Shao, C.F., 1993. Fuzzy mathematical programming for traffic safety planning on an urban expressway. Transportation Planning and Technology 17, pp. 179-190.

Akiyama, T., Shao, C-F. and Sasaki, T., 1994. Traffic Flow on urban networks with fuzzy information. Memorial Faculty of Engineering, Kyoto University 56, pp. 1-22.

Akiyama, T. and Tsuboi, H., 1996. Description of route choice behaviour by multi-stage fuzzy reasoning. Paper presented at the Highways to the Next Century Conference, Hong Kong.

Al-Deek, H. M, and Chandra, C. V. S. R., 2004. New algorithms for filtering and imputation of real-time and archived dual-loop detector data in I-4 data warehouse. Transportation Research Record 1867, pp.116-126.

American Society for Testing and Materials, ASTM, 1991. Standard Practice E1442, Highway Traffic Monitoring Standards, Philadelphia, PA.

Anderson, T.W. and Sclove, S.T., 1986. The statistical analysis of data. The Scientific Press, 2nd edn.

Bagchi, M. and White, P.R., 2005. The potential of public transport smart card data. Transport Policy, 12, pp. 464–474.

Barry, J.J., Newhouser, R., Rahbee, A., and Sayeda, S., 2002. Origin and destination estimation in New York City with automated fare system data. Transportation Research Record 1817, pp. 183–187.

Bellman, R. E., and Zadel, L. A., 1970, Decision Making in Fuzzy Environment. *Management Science*, Vol. 17, No. 4, pp. 141–164.

Blythe, P., 2004. Improving public transport ticketing through smart cards. *Proceedings of the Institute of Civil Engineers, Municipal Engineer 157*, pp. 47–54.

CEGASA. Available at <http://www.cegasatraffic.com/es/listado/b2/TomaDatos.html>. Last accessed: 30 September 2010.

Chakroborthy, P., Kikuchi, S., 1990. Application of fuzzy set theory to the analysis of capacity and level of service of highways. In: Ayyub, B.M. (Ed.), *Proceedings of ISUMA, The First International Symposium on Uncertainty Modeling and Analysis*. IEEE Computer Press, College Park, Maryland, pp. 146-150.

Chakroborty, P. and Kikuchi, S., 2003. Calibrating the membership functions of the fuzzy inference system: instantiated by car-following data. *Transportation Research Part C 11(2)*, pp. 91-119.

Chambers, R., 2001. Evaluation Criteria for Statistical Editing and Imputation. *National Statistics Methodological Series*, pp. 28- 41.

Chang, Y.-H. and Shyu, T.-H., 1993. Traffic signal installation by the expert system using fuzzy set theory for inexact reasoning. *Transportation Planning and Technology 17*, pp. 191-202.

Chen, L., May, A., Auslander, D., 1990. Freeway ramp control using fuzzy set theory for inexact reasoning. *Transportation Research 24A*, pp. 15-25.

Chen, H., 2002. Stochastic Optimization in Computing Multiple Headways for a Single Bus Line. *Proceedings of the 35th Annual Simulation Symposium IEEE*.

Chen, C., Kwon, J., Rice, J., Skabardonis, A. and Varaiya, P., 2003. Detecting errors and imputing missing data for single-loop surveillance systems. *Transportation Research Record 1855*, pp.160-167.

Chiu, S., 1992. Adaptive traffic signal control using fuzzy logic. Proceedings of the Intelligent Vehicles Symposium, Detroit, MI, pp. 98-107.

Coifman, B., 1999. Using dual loop speed traps to identify detector errors Transportation Research Record, 1683, pp. 47–58.

Delgado, M., Verdegay, J.L. and Vila, M.A. 1992. Linguistic Decision-Making Models. International Journal of Intelligent Systems, 7, 479-492.

Dempsey, S.P., 2008. Privacy Issues with the Use of Smart Cards. Legal Research Digest, p. 25.

Geng, Y. and Wu, X., 2008. The erroneous data imputation models for Beijing's urban traffic flow data by time series and correlation analysis. Transportation Research Board 87th Annual Meeting, Washington, D.C., January 13-17, on CD-ROM.

Gold, D. L., Turner, S. M., Gajewski, B. J. and Spiegelman, C., 2000. Imputing missing values in ITS data archives for intervals under 5 minutes. Presented at the 80th Annual Meeting of the Transportation Research Board, Washington, DC.

Homaifar, A. and McCormick, E., 1995. Simultaneous design of membership functions and rule sets for fuzzy controllers using genetic algorithms. IEEE Transactions on Fuzzy Systems 3 (1), pp. 129-139.

Kaczmarek, M., 2005. Fuzzy group model of traffic flow in street networks. Transportation Research Part C 13(2), pp. 93-105.

Kalić ,M. and Teodorović , D., 1996. Solving the trip distribution problem by fuzzy rules generated by learning from examples. Proceedings of the XXIII Yugoslav Symposium on Operations Research, Zlatibor, Yugoslavia, pp. 777-780 (in Serbian).

Kalić , M. and Teodorović , D., 1997a. Trip distribution modeling using soft computing techniques. Paper presented at the EURO XV/ INFORMS XXXIV (Book of abstracts, p. 74), Barcelona.

Kalić , M. and Teodorović , D., 1997b. A soft computing approach to trip generation modeling. Paper presented at the 9th Mini EURO Conference Fuzzy sets in Traffic and transport systems, Budva, Yugoslavia.

Kikuchi, S., 1997. Method to Defuzzify the Fuzzy Number: Transportation Problem Application. Presented at 1st Euro Conference on Fuzzy-Set Theory Application to Transportation, Budva, Yugoslavia.

Kikuchi, S. and Miljkovic, D., 1999. Method to Preprocess Observed Traffic Data for Consistency: Application of Fuzzy Optimization Concept. Transportation Research Record 1679, pp. 73-80.

Kikuchi, S., Miljkovic, D. and Van Zuylen, H.J., 2000. Examination of Methods that Adjust Observed Traffic Volumes on a Network. Transportation Research Record 1717, pp. 109-119.

Klir, G. J., and Wierman, M. J., 1999. Uncertainty-Based Information Elements of Generalized Information Theory. Physica-Verlag, Heidelberg, Germany.

Kwon, J., Chen, C. and Varaiya, P., 2004. Statistical Methods for Detecting Spatial Configuration Errors in Traffic Surveillance Sensors," Transportation Research Record no. 1870, Transportation Research Board, pp. 124-132.

Kwon, J., Petty, K., Shieh, E., Kopelias, P. and Papandreou, K., 2008. An automatic Method for Imputing and Balancing Link Traffic Counts. Transportation Research Board 87th Annual Meeting, Washington, D.C., January 13-17, on CD-ROM.

Laaksonen, S., 1999. How to find the best Imputation Technique? Test with various methods. International Conference on Survey Nonresponse. Portland, Oregon, pp. 28-31.

Larkin, L., 1985. A fuzzy logic controller for aircraft flight control. In: Sugeno, M. (Ed.), Industrial applications of fuzzy control. Elsevier Science (North-Holland), New York, pp. 87-103.

Lee, S., Krammes, R.A. and Yen, J., 1998. Fuzzy Logic-based incident detection for signalized diamond interchanges. *Transportation Research Part C* 6(5-6), pp. 359-377.

Lin, D.-Y., Boyles, S., Valsaraj, V. and Waller, S.T., 2012. Fuzzy reliability assessment for traffic data, *J. Chin. Inst. Eng.*, 35, (3), pp. 1–14.

LINPROG. Available at <http://www.mathworks.com/help/toolbox/optim/ug/linprog.html>. Last accessed: 22 March 2011.

Lotan, T. and Koutsopoulos, H., 1993a. Route choice in the presence of information using concepts from fuzzy control and approximate reasoning. *Transportation Planning and Technology* 17, pp. 113-126.

Lotan, T. and Koutsopoulos, H., 1993b. Models for route choice behaviour in the presence of information using concepts from fuzzy set theory and approximate reasoning. *Transportation* 20, pp. 129-155.

Lotfi, F. H., Allahviranloo, T., Jondabeh, M. A. and Alizadeh, L., 2009: Solving a full fuzzy linear programming using lexicography method and fuzzy approximate solution. *Applied Mathematical Modelling* 33, pp. 3151–3156.

Marzano, V., Papola, A. and Simonelli, F., 2008. Investigating the effectiveness of the O/D matrix correction procedure using traffic counts. *Transportation Research Board 87th Annual Meeting*, Washington, D.C., January 13-17, on CD-ROM.

Mendenhall, W. and Sincich, T., 1988. *Statistics for the engineering and computer sciences*. Dellen Publishing Company, San Francisco, CA, USA, 2nd edn.

Milosavljević, N., Teodorović, D., Papić, V. and Pavković, G., 1996. A fuzzy approach to the vehicle assignment problem. *Transportation Planning and Technology* 20, pp. 33-47.

Nakatsuyama, M., Nagahashi, N., Nishizuka, N., 1983. Fuzzy logic phase controller for traffic functions in the one-way arterial road. Proceedings IFAC 9th Triennial World Congress. Pergamon Press, Oxford, pp. 2865-2870.

Ndoh, N.N. and Ashford, N.J., 1994. Evaluation of transportation level of service using fuzzy sets. Paper presented at the 73rd Annual Meeting, Transportation Research Board, Washington, D.C.

Nguyen, L. H. and Scherer, W. T., 2003. Imputation Techniques to Account for Missing Data in Support of Intelligent Transportation Systems Applications. Technical report UVACTS-13-0-78, University of Virginia, Center for Transportation Studies.

Nihan, N.L. and Davis, G.A., 1987a. Application of prediction-error minimization and maximum likelihood to estimate intersection O/D matrices from traffic counts, *Transp. Sci.*, 23, pp. 77–90.

Nihan, N.L. and Davis, G.A., 1987b. Recursive estimation of origin-destination matrices from input/output counts, *Transp. Res. B*, 21, pp. 149–163.

Nihan, N.L. and Davis, G.A., 1989. Application of prediction-error minimization and maximum likelihood to estimate intersection O/D matrices from traffic counts, *Transp. Sci.*, 23, pp. 77–90.

Nihan, N.L., Jacobson, N., Bender, J. D. and Davis, G., 1990. Detector data validity Technical Report WA-RD 208.1, Washington State Transportation Center.

Nihan, N. L., Zhang, X. and Wang, Y., 2002. Evaluation of dual-loop error using video ground truth data. Technical Report Joint Report TNW02-02, WA-RD 535.1, Transportation Northwest/Washington State Department of Transportation.

Pappis, C. and Mamdani, W., 1977. A fuzzy logic controller for a traffic junction. *IEEE Transactions on Systems, Man, and Cybernetics SMC-7*, pp. 707-717.

Pattnaik, S.B. and Ramesh Kumar, K., 1996. Level of service of urban roads based on users' perception. *Civil Engineering Systems* 14, pp. 87-110.

Payne, H.J., Helfenbein, E.D. and Knobel, H.C., 1976. Development and testing of incident detection algorithms. Technical report FHWA-RD-76-20, Federal Highway Administration, US Department of Transportation.

Pelletier, M.P., Trépanier, M., and Morency, C., 2011. Smart card data use in public transit: a literature review. *Transportation Research Part C* 19, pp. 557–568.

Pentrice, G., 1987. Problems of Present Data Collection and Analysis. Proc. Conference of the Institution of Civil Engineers, Institution of Civil Engineers, London.

Perincherry, V. and Kikuchi, S., 1990. A fuzzy approach to the transshipment problem. In: Ayyub, B.M. (Ed.), *Proceedings of ISUMA, The First International Symposium on Uncertainty Modeling and Analysis*. IEEE Computer Press, College Park, Maryland, pp. 330-335.

Qing Z., Yingzhe W. and Jiankou L., 2009. Public Transport IC Card Data Analysis and Operation Strategy Research Based on Data Mining Technology. *International Forum on Computer Science-Technology and Applications*.

Quadrado, J.C. and Quadrado, A.F., 1996. Fuzzy modeling of accessibility: Case study Lisbon metropolitan area. *Proceedings of the Fourth European Congress on Intelligent Techniques and Soft Computing*, Aachen, Germany, pp. 1307-1311.

Rudy, K., Wang, H. and Ni, D., 2008. Modeling and Optimization of Link Traffic Flow. *Transportation Research Board 87th Annual Meeting*, Washington, D.C., January 13-17, on CD-ROM.

Saameño Rodriguez, J.J., Guerrero Garcia, C., Muñoz Pérez, J. and Mérida Casermeiro, E., 2006. A general model for the undesirable single facility location problem', *Oper. Res. Lett.*, 34, (4), pp. 427–436.

Sasaki, T. and Akiyama, T., 1986. Development of fuzzy traffic control system on urban expressway. Preprints 5th IFAC/IFIP/IFORS International Conference in Transportation Systems, pp. 333-338.

Sasaki, T. and Akiyama, T., 1987. Fuzzy on-ramp control model on urban expressway and its extension. In: Gartner N.H., Wilson, N.H.M. (Eds), Transportation and traffic theory. Elsevier Science, New York, pp. 377-395.

Sasaki, T. and Akiyama, T., 1988. Traffic control process of expressway by fuzzy logic. Fuzzy Sets and Systems 26, pp. 165-178.

Sayed, T., Abdelwahab, W. and Navin, F., 1995. Identifying accident-prone locations using fuzzy pattern recognition. Journal of Transportation Engineering 121, pp. 352-358.

Sharma, S.C., Kilburn, P. and Wu, Y.Q., 1996. The precision of AADT volumes estimates from seasonal traffic counts. Alberta example. Canadian Journal of Civil Engineering 23 (1), pp. 302-304.

Schretter, N. and Hollatz, J., 1996. A fuzzy logic expert system for determining the required waiting period after traffic accidents. Proceedings of the Fourth European Congress on Intelligent Techniques and Soft Computing, Aachen, Germany, pp. 2164-2170.

Silva-Ramírez, E.L., 2007. Redes de Neuronas Artificiales en la edición e imputación de datos. Ph.D Thesis. University of Sevilla, (in Spanish).

Smith, P.N., 1993. Fuzzy evaluation of potential suburban railway station locations. Journal of Advanced Transportation 27, pp. 153-179.

Srinivasan, D., Sanyal, S. and Sharma, V., 2007. Freeway incident detection using hybrid fuzzy neural network, IET Intell. Transp. Syst., 1, (4), pp. 249–259.

Tanaka, H., 1990. Fuzzy Modeling and Its Applications. Asakura Shoten, Tokyo, Japan.

Tang, S. and Gao, H., 2005. Traffic-incident detection-algorithm based on nonparametric regression, *IEEE Trans. Intell. Transp. Syst.*, 6, (1), pp. 38–42.

Tavana, H. and Mahmassani, H., 2000. Estimation of dynamic origin-destination flows from sensor data using bi-level optimization method. *Proc. 80th Annual Meeting of the Transportation Research Board*, CD ROM.

Teodorović, D. and Babić, O., 1993. Fuzzy Inference Approach to the Flow Management Problem in Air Traffic Control. *Transportation Planning and Technology* 17, pp. 165 - 178.

Teodorović, D. and Kalić, M., 1996. Solving the modal split problem by fuzzy rules generated by learning from examples. *Proceedings Information Technologies, Zabljak, Yugoslavia*, pp. 48-54 (in Serbian).

Teodorović, D. and Kikuchi, S., 1991. Application of fuzzy sets theory to the saving based vehicle routing algorithm, *Civil Engineering Systems*, 8, pp. 87 - 93.

Teodorović, D., Kalić, M. and Pavković, G., 1994. The Potential for Using Fuzzy Set Theory in Airline Network Design, *Transportation Research*, 28B, pp. 103-121.

Thomas, T. and van Berkum, E.C., 2009. Detection of incidents and events in urban networks', *IET Intell. Transp. Syst.*, 3, (2), pp. 198–205.

Trépanier, M., Tranchant, N. and Chapleau, R., 2007. Individual trip destination estimation in a transit smart card automated fare collection system. *Journal of Intelligent Transportation Systems* 11, pp. 1–14.

Turner, S., Albert, L., Gajewski, B., and Eisele, W., 2000. Archived intelligent transportation system data quality: preliminary analyses of San Antonio TransGuide data. *Transportation Research Record* 1719, pp.77-84.

Tussel, F., 2002. Neural Networks and Predictive Matching for Flexible Imputation. *DataClean 2002 Conference*. Jyväskylä (Finland), pp. 29-31.

Tzeng, G.H. and Teng, J.-Y., 1993. Transportation investment project selection with fuzzy multi-objectives. *Transportation Planning and Technology* 17, pp. 91-112.

Tzeu-Chen Han, 2008. Application of Fuzzy Regression on Air Cargo Volume Forecast. Transportation Research Board 87th Annual Meeting, Washington, D.C., January 13-17, on CD-ROM.

Vanajahshi, L. and Rillet, L.R., 2004. Loop detector data diagnostics based on conservation-of vehicles principle, *Transp. Res Rec.*, 1870.

Von Altrock, C., 1995. *Fuzzy logic and NeuroFuzzy applications explained*. Upper Saddle River, NJ: Prentice Hall PTR.

Vukadinović, K. and Teodorović, D., 1994. A Fuzzy Approach to the Vessel Dispatching Problem, *European Journal of Operational Research*, 76, pp. 155 - 164.

Wall, Z.R. and Dailey, D.J., 2003: Algorithm for detecting and correcting errors in archived traffic data, *Transp. Res. Rec.*, 1855, pp. 183–190.

Wang, L.-X. and Mendel, J., 1992a. Back propagation of fuzzy systems as nonlinear dynamic system identifiers. *Proceedings IEEE International Conference on Fuzzy Systems*, San Diego, pp. 807-813.

Wang, L.-X. and Mendel, J., 1992b. Fuzzy basis functions, universal approximation, and orthogonal least squares learning. *IEEE Transactions on Neural Networks* 3, pp. 807-813.

Wang, L.-X. and Mendel, J., 1992c. Generating fuzzy rules by learning from examples. *IEEE Transactions on systems, Man and Cybernetics* 22, pp. 1414-1427.

Xu, W. and Chan, Y., 1993a. Estimating an origin-destination matrix with fuzzy weights. Part 1: Methodology. *Transportation Planning and Technology* 17, pp. 127-144.

Xu, W. and Chan, Y., 1993b. Estimating an origin-destination matrix with fuzzy weights Part 2: Case studies. *Transportation Planning and Technology* 17, pp. 145-164.

Yager, R., and Filev, D. P., 1994. *Essentials of Fuzzy Modeling and Control*. John Wiley and Sons, New York.

Zadeh, L.A., 1965. Fuzzy sets, *Information and Control*, 8 (3), pp. 338-353.

Zadeh, L., 1996. Fuzzy logic=computing with words. *IEEE Transactions on Fuzzy Systems* 4, 103-111.

Zhang, H.-Z., Wang, J. and Zi-hui Ren, Z.-H., 2008. Rough sets and FCM-based neuro-fuzzy inference system for traffic incident detection. *ICNC'08. Fourth Int. Conf. on Natural Computation*, vol. 7, pp. 260–264.

Zhao, J., Rahbee, A. and Wilson, N., 2007. Estimating a rail passenger trip origin–destination matrix using automatic data collection systems. *Computer-Aided Civil and Infrastructure Engineering* 22, pp. 376–387.

Zhong, M. and Sharma, S., 2003. Development of Improved Models for Imputing Missing Traffic Counts. *The Open Transportation Journal*, 3, pp. 35-48.

Zhong, M., Lingras, P. and Sharma, S., 2004. Estimation of missing counts using factor, genetic, neural, and regression techniques. *Transportation Research Part C* 12(2), pp. 139-166.

Zimmermann, H. J., 1996, *Fuzzy Set Theory and Its Applications*, 3rd ed. Kluwer Academic Publishers, Norwell, Mass.

Zimmermann, H. J., 2001. *Fuzzy Set Theory and Its Applications*, 4th ed. Springer, Berlin.

APPENDIX: PUBLISHED PAPERS



Contents lists available at ScienceDirect

Transportation Research Part C

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

Bilevel fuzzy optimization to pre-process traffic data to satisfy the law of flow conservation

J. de Oña^{a,*}, P. Gómez^b, E. Mérida-Casermeyro^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, ETS Computer Science Engineering, Boulevard Louis Pasteur s/n, 29071 Málaga, Spain

ARTICLE INFO

Article history:

Received 5 April 2008

Received in revised form 12 February 2010

Accepted 16 February 2010

Keywords:

Traffic counts

Fuzzy logic

Transport planning

Optimization

Data consistency

Subjective analyst knowledge

ABSTRACT

Traffic data obtained in the field usually have some errors. For instance, traffic volume data on the various links of a network must be consistent and satisfy flow conservation, but this rarely occurs. This paper presents a method for using fuzzy optimization to adjust observed values so they meet flow conservation equations and any consistency requirements. The novelty lies in the possibility of obtaining the best combination of adjusted values, thereby preserving data integrity as much as possible. The proposed method allows analysts to manage field data reliability by assigning different ranges to each observed value. The paper is divided into two sections: the first section explains the theory through a simple example of a case in which the data is equally reliable and a case in which the observed data comes from more or less reliable sources, and the second one is an actual application of the method in a freeway network in southern Spain where data were available but some data were missing.

© 2010 Elsevier Ltd. All rights reserved.

1. Introduction

The study of any transport system requires enormous quantities of data and an understanding of their dependence on each other. Arguably, volume is the most important traffic datum of them all. Field data is generally inconsistent, and therefore they need to be processed in a way that will make them consistent before they can be used in algorithms for prediction, monitoring and decision-making purposes. The methods used to estimate Origin–Destination (O/D) matrices are based on the hypothetical availability of precise traffic volume data and reliable preliminary O/D data. The input data for most traffic networks, however, are either unavailable or contain measurement errors, as in the case of traffic counts and sensor speed measurements. In fact, some studies (Zhong et al., 2004) demonstrate that 50% of the Permanent Traffic Counts (PTCs) set up on highways contain lost data, making it difficult to ignore measurement errors when processing data used to plan, design, control and manage traffic (Sharma et al., 1996). The existence of errors makes data obtained in the field difficult to manage and to analyze.

In the past, certain methods were applied to adjust the observed values so they would comply with flow conservation laws at each network node, aside from other requirements that values need to meet before they can be used as input data in traffic planning algorithms. The methods used were manual value adjustment, least square adjustment and the maximum likelihood method (Kikuchi et al., 2000). Recently, new methods of value adjustment based on fuzzy logic have been developed to preserve data integrity as much as possible. The methods are: fuzzy regression, fuzzy optimization and necessity-interval-regression method (Kikuchi et al., 2000). A number of important publications on fuzzy logic have been submitted

* 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-Casermeyro).

over the past 20 years, although most of them are based on the fields of deduction and control in situations of complex behavior. Pappis and Mamdani (1977) were the first to apply fuzzy logic to transport; specifically, to traffic signal controllers.

Lost data processing is another frequent issue. When available input data exist at all, they often contain errors due to the sensors' operating faults (Kwon et al., 2008). From a formal viewpoint, the problem of debugging input data in order to avoid inconsistency and of assigning values to missing data has generally been analyzed by an area of Statistics (Data Editing and Imputation). Most efforts have focused on processing 'missing values', and on detecting and debugging. Inconsistencies have been avoided by using redundant or related information. Some classical techniques are: imputation by mean, median, regression or hot-deck (Chambers, 2001; Laaksonen, 1999). Recently, some new techniques based on Artificial Intelligence and on neural networks, in particular, are being developed (Silva-Ramírez, 2007; Tussel, 2002). Certain authors (Kaczmarek, 2005; Marzano et al., 2008; Rudy et al., 2008) have submitted methods based on the characteristics of erroneous traffic data in urban networks, supplemented with the latest data imputation models (Lee et al., 1998; Geng and Wu, 2008). Other methods based on weighted least squares regression also exist, such as the methods submitted by Kwon et al. (2008).

The aim of this article is to submit a method whereby field data could be pre-processed to make them consistent while preserving their integrity as much as possible, and which would include their reliability as perceived subjectively by the analyst. The method is based on fuzzy logic and is intended to optimize the solution obtained. The result would be a reliable solution that comes close to the observed values, thereby resolving measurement errors in traffic counts. The method also allows field data to be processed when there are lost values.

2. Description of the problem

A simple freeway network is used to explain the method. Consider the situation shown in Fig. 1, in which real consistent data are available (Table 1, column 2). The data are used to simulate a scenario with non consistent data: traffic counts from the database are randomized within ±25% of their values at all intersections to simulate a case in which data is not consistent (Table 1, column 3). Next, the randomly obtained data in the database are considered to be field data; i.e. the observed values (OV).

Theoretically, in any transport network such as the one shown in Fig. 1, the total "incoming volumes" should be equal to the total "outgoing volumes" at any node in the network and in any flow direction in such a way that the law of conservation of flow is satisfied. In the simulated scenario, (Table 1, column 3), however, it is found that:

$$\begin{aligned}
 x_3 + x_5 &\neq y_5 + y_6 \\
 x_1 + x_2 &\neq y_1 + y_4 \\
 y_1 + y_2 &\neq z_1 + z_4 \\
 y_3 + y_5 &\neq z_5 + z_6 \\
 y_2 + y_6 &\neq w_{10} + w_{11} + w_{12} \\
 y_3 + y_4 &\neq w_1 + w_5 + w_7
 \end{aligned}$$

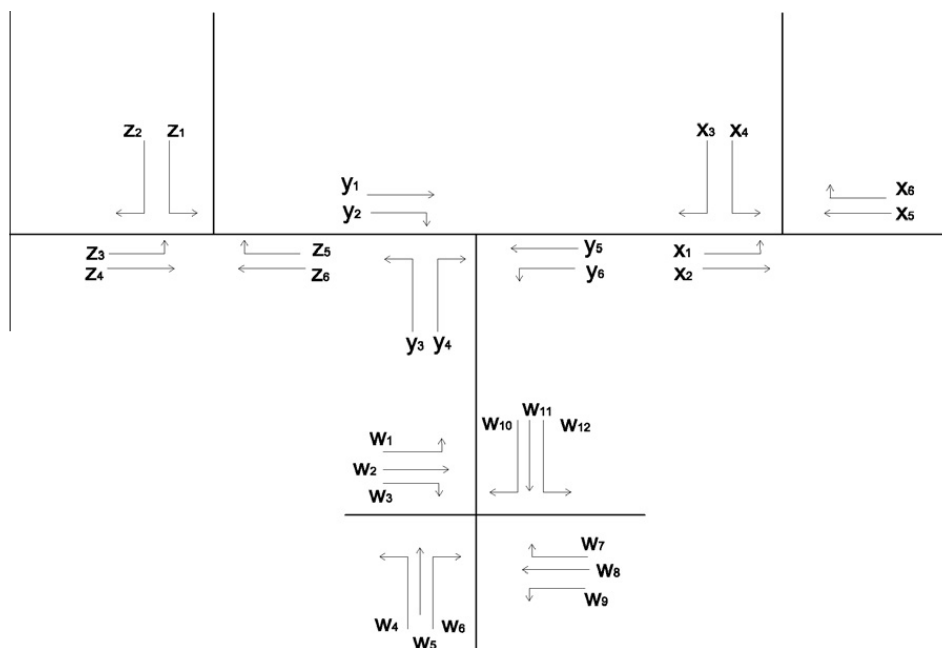


Fig. 1. Simple freeway network used to explain the method.

Table 1

Example 1: base data, randomized inconsistent data, adjusted data, and results for different α ranges.

(1)	(2)	(3)	(4)			(5)			(6)			(7)			(8)			(9)				
			RV	OV	MM method		$\alpha = 0.4$	MM method		α^a	MS method		$\alpha = 0.4$	MS method		α^a	BO method		$\alpha = 0.4$	BO method		α^a
					AV	h_i		Δ	AV		h_i	Δ		AV	h_i		Δ	AV		h_i	Δ	
w_1	135	113	128	0.68	7	126	0.63	9	113	1.00	22	113	1.00	22	128	0.68	7	126	0.63	9		
w_2	30	37	39	0.87	9	40	0.91	10	37	1.00	7	37	1.00	7	37	1.00	7	37	1.00	7		
w_3	43	42	42	1.00	1	44	0.92	1	42	1.00	1	42	1.00	1	42	1.00	1	42	1.00	1		
w_4	104	96	95	0.97	9	97	0.97	7	96	1.00	8	96	1.00	8	95	0.97	9	97	0.97	7		
w_5	148	134	152	0.67	4	150	0.61	2	134	1.00	14	134	1.00	14	152	0.67	4	150	0.61	2		
w_6	19	18	21	0.59	2	20	0.82	1	18	1.00	1	18	1.00	1	18	1.00	1	18	1.00	1		
w_7	28	27	30	0.73	2	30	0.64	2	27	1.00	1	27	1.00	1	30	0.73	2	30	0.64	2		
w_8	35	37	40	0.80	5	40	0.91	5	37	1.00	2	37	1.00	2	37	1.00	2	37	1.00	2		
w_9	22	18	21	0.59	1	20	0.82	2	18	1.00	4	18	1.00	4	18	1.00	4	18	1.00	4		
w_{10}	102	78	77	0.97	25	79	0.96	23	78	1.00	24	78	1.00	24	77	0.97	25	79	0.96	23		
w_{11}	175	171	172	0.99	3	169	0.96	6	171	1.00	4	171	1.00	4	172	0.99	3	170	0.98	5		
w_{12}	3	4	4	1.00	1	4	1.00	1	4	1.00	1	4	1.00	1	4	1.00	1	4	1.00	1		
x_1	265	215	253	0.57	12	255	0.71	10	220	0.94	45	220	0.96	45	253	0.57	12	253	0.73	12		
x_2	54	53	62	0.59	8	61	0.77	7	53	1.00	1	53	1.00	1	62	0.59	8	61	0.77	7		
x_3	105	116	109	0.85	4	111	0.93	6	116	1.00	11	116	1.00	11	109	0.85	4	111	0.93	6		
x_4	110	132	130	0.96	20	133	0.99	23	132	1.00	22	132	1.00	22	130	0.96	20	133	0.99	23		
x_5	200	177	168	0.88	32	168	0.92	32	161	0.78	39	161	0.86	39	168	0.88	32	168	0.92	32		
x_6	58	51	48	0.86	10	52	0.97	6	51	1.00	7	51	1.00	7	50	0.95	8	51	1.00	7		
y_1	26	31	26	0.61	0	28	0.79	2	20	0.13	6	20	0.25	6	26	0.61	0	26	0.66	0		
y_2	20	17	15	0.71	5	14	0.62	6	17	1.00	3	17	1.00	3	15	0.71	5	15	0.75	5		
y_3	18	21	21	1.00	3	18	0.70	0	21	1.00	3	21	1.00	3	21	1.00	3	18	0.70	0		
y_4	293	353	289	0.56	4	288	0.61	5	253	0.31	40	253	0.40	40	289	0.56	4	288	0.61	5		
y_5	45	39	39	1.00	6	41	0.89	4	41	0.87	4	41	0.89	4	39	1.00	6	41	0.89	4		
y_6	260	226	238	0.87	22	238	0.89	22	236	0.89	24	236	0.91	24	238	0.87	22	238	0.89	22		
z_1	33	26	29	0.72	4	29	0.75	4	26	1.00	7	26	1.00	7	29	0.72	4	29	0.75	4		
z_2	22	17	20	0.57	2	20	0.87	2	17	1.00	5	17	1.00	5	17	1.00	5	17	1.00	5		
z_3	25	27	29	0.82	4	31	0.92	6	27	1.00	2	27	1.00	2	27	1.00	2	27	1.00	2		
z_4	13	11	12	0.78	1	13	0.61	0	11	1.00	2	11	1.00	2	12	0.78	1	12	0.81	1		
z_5	28	33	32	0.93	4	31	0.87	3	33	1.00	5	33	1.00	5	32	0.93	4	31	0.87	3		
z_6	35	29	28	0.92	7	28	0.93	7	29	1.00	6	29	1.00	6	28	0.92	7	28	0.93	7		
g = sum			24.04			24.92			27.93			28.26			25.89			25.98				
h_i																						
h = min			0.56			0.61			0.13			0.25			0.56			0.61				
h_i																						
Average			7.23			7.13			10.70			10.70			7.10			6.97				
Δ																						

Note: RV (real value); OV (observed value); AV (adjusted value); Δ (difference between RV and AV in absolute value).

^a $\alpha = 0.65$ for x_i ; $\alpha = 0.5$ for y_i ; and z_i ; $\alpha = 0.3$ for w_i .

Actually, this is usually the case, particularly when the network is large. Pentrice (1987) stated that data inconsistency is inevitable even in a well-controlled survey, but volume count consistency at different links is critical to ensuring the integrity of the results of any of the ITS-related algorithms.

When the network becomes larger, the possibility of inconsistency in traffic volume counts increases, so flow conservation is more difficult. The concern in this paper is how to adjust the individual observed volumes to a set of new values that satisfy the flow conservation principle at any point in the network. Furthermore, the adjustment should be such that the integrity of the observed values is preserved as much as possible. To this end, a fuzzy optimization method is used to obtain adjusted values that comply with the law of flow conservation and that resemble consistent real data as closely as possible. In this example, the integrity of the results obtained can be verified with the available real consistent data.

3. The bilevel fuzzy optimization method

The search for the “best” set of adjusted values is an optimization process that aims to find a set of values close to the observed ones that verifies the conservation of flow principle.

The proposed method is based on the following concept: each observed value is considered an approximate value represented by a fuzzy number, defined by a membership function. If the value is x , it is interpreted as “approximately x ”. The true value is considered to lie near x . The method attempts to find an adjusted value as close to the observed value as possible while satisfying the conservation of flow at every point in the network. This is accomplished by applying the concept of fuzzy optimization developed in fuzzy set theory.

Given a set of observed values, there are an infinite number of combinations of adjusted values, each of which satisfies the set of flow conservation equations. For a given combination, the membership grade $h_{x_i}(x'_i)$ of each adjusted value (x'_i) in the corresponding fuzzy set (x_i) is calculated. Three methods of optimization could be used:

- a. by maximizing the minimum $h_{x_i}(x'_i)$ for all i ,
- b. by maximizing the sum of $h_{x_i}(x'_i)$, and,
- c. by maximizing the minimum $h_{x_i}(x'_i)$ for all i at one level and, after this has been achieved, by applying a second level of optimization by maximizing the sum of $h_{x_i}(x'_i)$. Thus, the combination with the highest sum is selected from among all the combinations that could maximize the lowest membership grade. The value with the least membership grade is taken into consideration, and also all the other observed data.

In case (a) (MM method), the lowest membership grade for the combination is recorded. By comparing the lowest membership grades among all the combinations of traffic volumes, the one that has the highest value is chosen as the best combination of a set of adjusted values. This method was already introduced by Kikuchi and Miljkovic (1999).

On the other hand, in the objective function sum of $h_{x_i}(x'_i)$'s (case (b)) (MS method) for a given combination, the membership grade of each adjusted value in the corresponding fuzzy set is calculated. The sum of the membership grades among all the combinations of traffic volumes is recorded, and the one that has the highest sum of membership grades is chosen as the best combination of a set of adjusted values.

The third possibility is a two step way of optimization or bilevel optimization method (BO method). In step one, case (a), the lowest membership grade is maximized. In step two, the membership grades that would produce the largest possible $\max(\min(h_i))$ and that would seek to increase the value of all of the h_i at the same time (which would achieve the sum of both) are summed up and maximized.

The MM method can attend to a set of data which its minimum membership grade is maximized but the problem is that an infinite number of combinations could satisfy this condition and the MM method randomly chooses one of them. The BO method chooses a set that while it satisfies that condition; it optimizes the rest of the values, maximizing the membership grade of all the data, so the BO method uses both ways of optimization in order to improve the solution.

The mathematical steps involved in addressing the optimization problem are:

1. Use fuzzy numbers to represent observed values.
2. Formulate the objective and constraints.
3. Solve as a mixed linear programming problem.

The process is explained step by using the simple highway network shown in Fig. 1.

3.1. Using fuzzy numbers to represent observed values

The observed values are “fuzzified” and are considered a fuzzy set with a triangular membership function.

Fig. 2 shows the shape of the membership function with the centre value x_i and a range $[x_i - \alpha x_i, x_i + \alpha x_i]$, where α is a constant higher than 0. The triangular membership function is not a prerequisite but, in the absence of any other information, this is a reasonable assumption, and such assumption is often used in fuzzy set theory (Zimmermann, 2001).

The selection of the constant α depends on the judgement of the analyst with respect to the adjusted value's acceptable deviation from the observed value. This value allows the analyst to enter the reliability of each datum (i.e. the more reliable data will have a lower value of α than if they were less reliable). If only one value of α is used for all data, the scope of the range has little effect on the final adjusted values, once it is broad enough for a feasible set of solutions to be found.

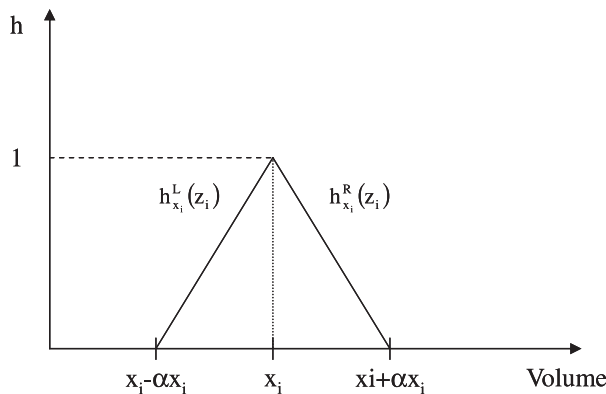


Fig. 2. Triangular membership function.

The membership function is defined for the left- and right-hand sides of the triangle. For an observed value of x_i and the assumed range $[x_i - \alpha x_i, x_i + \alpha x_i]$, the general expression of the membership functions is:

$$h_{x_i}(x'_i) = \begin{cases} h_{x_i}^L = \frac{x'_i - (x_i - \alpha x_i)}{\alpha x_i} & \text{if } x_i - \alpha x_i < x'_i \leq x_i \\ h_{x_i}^R = \frac{x'_i - (x_i + \alpha x_i)}{-\alpha x_i} & \text{if } x_i < x'_i \leq x_i + \alpha x_i \end{cases} \quad (1)$$

In this formula $-\infty < x_i - \alpha x_i \leq x_i \leq x_i + \alpha x_i < \infty$, the triangular fuzzy number x_i is presented by $(x_i - \alpha x_i, x_i, x_i + \alpha x_i)$.

For the sake of simplicity, a symmetric triangle is used in this paper for the membership function. However, the left and right-hand limits can be set separately. To solve this example problem, it is assumed that the value x_i is the observed value and that the value of $\alpha > 0$. So the value of α is the spread of triangular fuzzy number x_i . The narrower the spread area, the less fuzzy the evaluation data will be, hence more precise. To the contrary, fuzziness is higher and thus more vague and ambiguous when the spread area increases (Tzeu-Chen Han, 2008).

Some authors have researched calibration of the membership function extensively. The classical approach to calibration has been the intuitive trial and error process, in which the analyst modifies the shapes of the membership functions little by little until the predicted output approximately fits the output data obtained from the real world (Chakroborty and Kikuchi, 2003). However, this process is time consuming. Other authors have developed a systematic way of carrying out the trial and error process (Wang and Mendel, 1992a,b,c; Homaifar and McCormick, 1995). The purpose of calibration is to modify the membership functions of the Fuzzy Inference System (FIS) so that the outcome predicted by the model is equal (or nearly equal) to the outcome obtained in the real world. Therefore, Chakroborty and Kikuchi (2003) presented a method in which a representation framework allows the FIS parameters to be modified in relation to the bases. FIS outputs are dictated by the parameters that define the membership functions of the fuzzy sets appearing in the antecedents and the consequents of the rules and the algebraic operators used for the logical connectives and to determine the final inferred value. They have developed a procedure that calibrates the membership function of the fuzzy sets by transforming the inference system into an Artificial Neural Network format. They have applied this procedure to the complex control task of car-following, but this procedure has not been applied yet to an urban transport system or a large-scale civil infrastructure system.

3.2. Formulating the objective function and its constraints

In a fuzzy number representation of observed values, fuzzy optimization techniques would be used to search for the adjusted values. The mathematical formulation of the three proposed methods used to solve the problem would be as follows:

3.2.1. MM method

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

Subject to

- Constraints related to the membership functions:

$$h_{x_i}^L(x'_i) \geq h \quad h_{x_i}^R(x'_i) \geq h \quad h_i \geq h \quad \text{for } i = 1, k \quad (3)$$

which means there are $2k + k$ constraints (where k is the number of control points)

- Constraints related to the conservation of flow at each control point. The constraints are defined by reviewing the flow pattern at each node in Fig. 1 as follows:

$$\begin{aligned} x'_3 + x'_5 &= y'_5 + y'_6 \\ x'_1 + x'_2 &= y'_1 + y'_4 \\ y'_1 + y'_2 &= z'_1 + z'_4 \\ y'_3 + y'_5 &= z'_5 + z'_6 \\ y'_2 + y'_6 &= w'_{10} + w'_{11} + w'_{12} \\ y'_3 + y'_4 &= w'_1 + w'_5 + w'_7 \\ x'_i, y'_i, z'_i, w'_i &\geq \text{for all } i \end{aligned} \quad (4)$$

where x'_i, y'_i, z'_i, w'_i is the integer unknown adjusted values; x_i, y_i, z_i, w_i the fuzzy set corresponding to the observed value x_i ; $h_{x_i}(x'_i)$ the membership grade of x'_i in the fuzzy set x_i , the same treatment for y_i, z_i and w_i ; h is an operational parameter that represents the smallest membership grade among all $h_{x_i}(x'_i)$'s. Where $h_{x_i}^L(x'_i) \geq h$ and $h_{x_i}^R(x'_i) \geq h$, respectively, show the expressions for the left- and right-hand sides of the triangle.

3.2.2. MS method

$$\max(g) \text{ where } g \text{ is sum}(h_i) \quad (5)$$

Subject to the same constraints as in the MM method, with regard to the membership functions Eq. (3) and to the conservation of flow at each control point Eq. (4).

3.2.3. BO method

Step 1: The problem is solved using the MM method Eq. (2), and we obtain a value of $h = h^*$.

Step 2: The problem is solved using the MS method Eq. (5) subject to the same constraints with regard to the conservation of flow at each control point Eq. (4) as in the MM or MS method, and to the following constraints related to the membership functions:

$$h_{x_i}^L(x'_i) \geq h^* \quad h_{x_i}^R(x'_i) \geq h^* \quad h_i \geq h^* \quad \text{for } i = 1, k \quad (6)$$

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

If only $\max(h)$ is performed (MS method), there may be several imputations for the observed data that produce the same value for h (Tussel, 2002; Silva-Ramírez, 2007). Therefore, they would be the same from the objective function point of view, whereas, in fact, some are better than others. The combination (0.9, 0.9, 0.9), for instance, would have the same value as (0.9, 1, 1), whereas the latter is better than the former. On the other hand, if the objective function were just $\max(g)$ (MS method), some values would show a $h_i = 0.00$, despite the fact that almost all the rest are 1.00, which is of no interest. The bilevel optimization process (BO method) allows the combination where the remaining membership degrees are the highest ones to be chosen from among all the combinations where the lowest value of h is maximized.

3.3. Solving as a mixed linear programming problem

Since every x'_i must be an integer number and h_i are real numbers, this is a mixed linear programming formulation. A mixed linear programming algorithm is formulated for the problem to maximize the membership grade of the adjusted values.

In Fig. 1, the mixed linear programming algorithm consists of 90 (3×30 observed volumes) inequality constraints related to membership functions and six equations related to flow conservation.

3.4. Introduction of data reliability

The selection of the value of α depends on the judgement of the analyst with respect to the adjusted value's acceptable deviation from the observed value.

In a complex transport network, there may be permanent traffic count stations where count data are fairly reliable, and other nodes where counting is sporadic, as well as points where traffic volumes have not been measured. Therefore, to define the α parameter coherently, the method must allow the analyst to assign different values to the α parameter in order to define the membership functions of each observed value. The values will depend on whether the parameter belongs to a set of data that are highly reliable (permanent traffic count station), averagely reliable (sporadic count) or highly unreliable (lost data).

3.5. Example network

As shown in Fig. 1, the example consists in analysing a network of four intersections, of which three have six movements and one has twelve.

In this example, the real consistent data are known (RV) (Table 1 column 2). The data are used to simulate a scenario with non consistent data. The simulated data are considered the OV (Table 1 column 3).

In this example, it is considered that traffic count station W is a permanent station, so the values have maximum reliability and their α parameter is the lowest, $\alpha = 0.3$. The reliability of stations Y and Z is lower so α takes a value of 0.5 (sporadic count stations) and, finally, the data from traffic count station X is supposed to be the least reliable one, so α is assigned a value of 0.65.

3.6. Results

In this case, since real data were available, three indicators could be used to verify the goodness of the adjustment of each one of the three optimization methods used (MM, MS and BO methods):

- The first indicator is the lowest value of h , which indicates the membership grade of the worst adjusted value (the degree of compatibility between the adjusted value and the observed value). If the value of h is near zero, then the adjusted value is close to the right or left end of the base of the membership function; if the value of h is near 1, then the adjusted value is

close to the observed value. Therefore, the solution where the lowest value of h is maximum is chosen as the best solution from the point of view of this parameter.

- The second indicator is the sum of h_i . The best solution is where the sum of h_i is maximum, because the adjusted values are closer to the observed values and integrity is more preserved.
- The last indicator, for which the consistent real data are available, is the average of the differences between the real consistent values and the adjusted values.

The results for the three methods are given in Table 1, where the adjusted values (AV) and the value of the membership grade (h) for each observed value are shown. The membership grade of the individual AV is computed by entering the adjusted value (x'_i) in the respective membership function, $h_{x_i}(x'_i)$. The table also shows the effect of using different α values, depending on the reliability of the observed volumes at each intersection.

Column 1 of Table 1 shows each movement in nodes W, X, Y and Z. Column 2 shows the consistent RV used to obtain the OV that show inconsistencies by randomizing the values within $\pm 25\%$.

Columns 4, 6 and 8 in Table 1 shows the AV, the corresponding values of h (h_i) and the difference (Δ) between RV and AV in absolute value, using the MM method, MS method and BO method respectively for an α parameter of 0.4 in all cases:

- MM method's results are shown in column 4. The lowest value of h in column 4 ($h = 0.56$) indicates the membership grade of the worst adjusted value. In this case Σh_i is 24.04.
- In column 6, MS method's results show that whereas most of the adjusted values get $h = 1.00$, other values show lower h and h could even be 0.00, in order to manage the highest Σh_i . The lowest value of h is reached for y_1 ($h = 0.13$). This situation, therefore, is not desirable either, since it allows a set of values with some h very close to 0.00 to be considered, providing the sum is the maximum. In this case, Σh_i is 27.93.
- The BO method's results are shown in column 8. If columns 4 and 8 are compared, it can be seen that the minimum value of h remains the same ($h = 0.56$). However, there has been an increase in Σh_i , which has gone from 24.04 (MM method) to 25.89 (BO method). Thus, this new method allows a combination where the remaining membership degrees are the highest ones to be chosen from among all the combinations with the lowest value of h .

As explained above, introducing the analyst's knowledge of the different precisions of the data he is working with improves the results of the adjustment. This is shown in columns 5, 7 and 9 in Table 1 where the AV, h_i and Δ are calculated, using the three methods for different α parameters depending on the reliability of the data. The α values used in this example have been 0.65 for "X", 0.5 for "Y" and "Z" and 0.3 for "W".

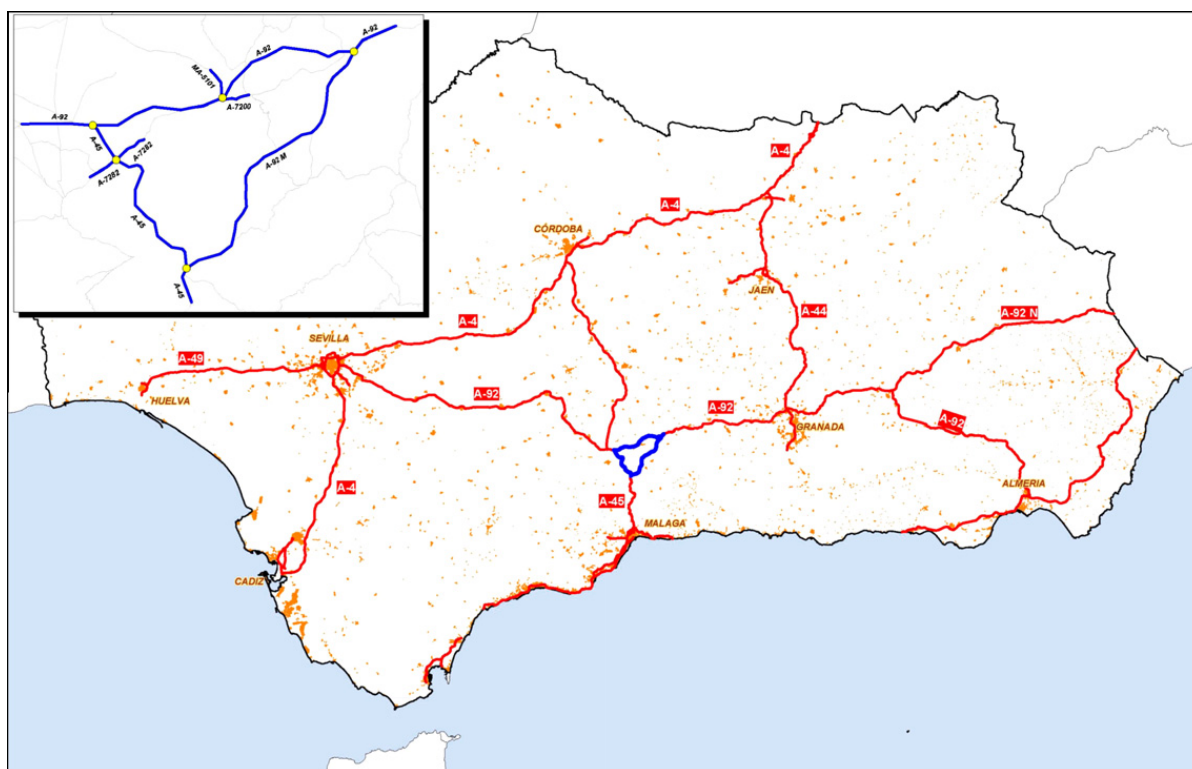


Fig. 3. Real intersections in Andalusia's freeway network (South of Spain).

As in the case of the same α for every observed value, for any α parameter, the MM and the BO methods obtain the same and a higher value of h minimum ($h = 0.61$) than the MS method ($h = 0.25$). However, the latter method obtains a higher value of Σh_i (28.26 versus 24.92 for the MM method and 25.98 for the BO method).

The results shown in column 5 are better than those shown in column 4. This is because the minimum value of h and Σh_i were higher and the value of average Δ was lower. Similar results are obtained by comparing columns 6–7 and 8–9 for the MS and the BO methods. This confirms the advantage of distinguishing between reliable data and less reliable data or, in other words, of introducing the subjective perception of the analyst.

The last row in Table 1 shows the average of Δ for each of the three methods used. It can be seen that the lowest value (6.97) is obtained for the BO method with different values of α , in comparison to the values of the MM method (7.13) and the MS method (10.70) with different values of α . This shows that the AV obtained with the BO method are closer to the real values than with the other two methods, so this is the method that best preserves the integrity of data.

4. Real intersections in Andalucía motorways network

Next, the three methods are used to adjust the traffic volumes of a series of adjacent intersections in Andalusia's freeway network (see Figs. 3 and 4) for which real and therefore inconsistent data are available. In this example, the parameter Δ is

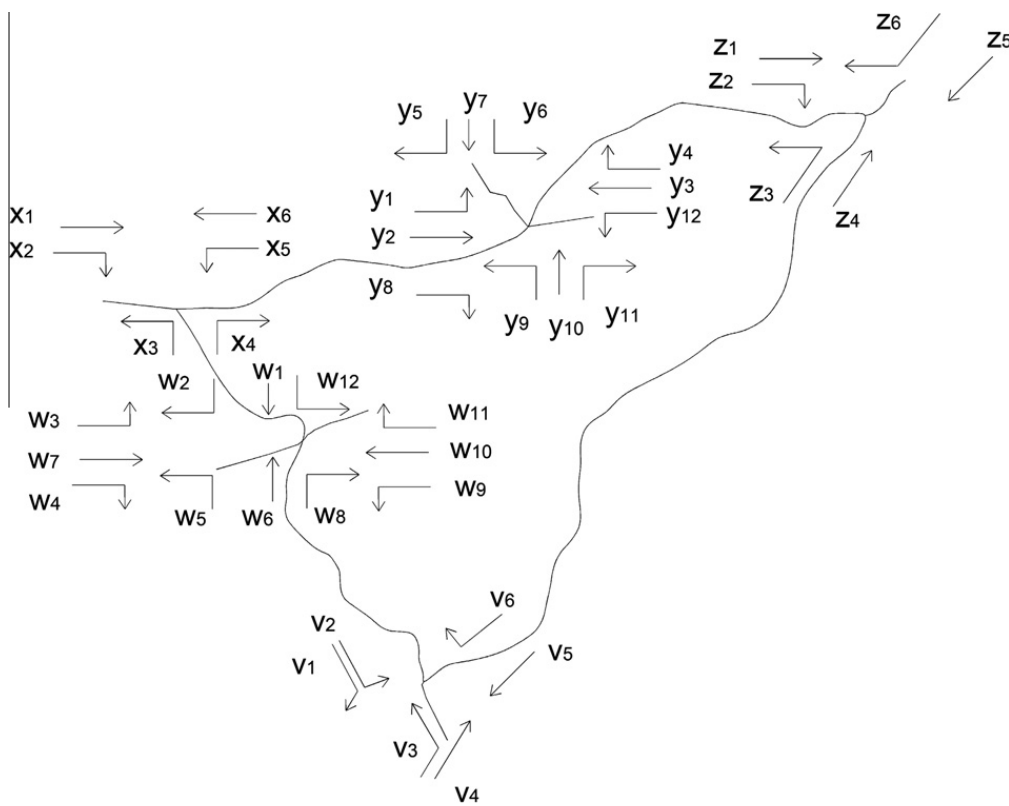


Fig. 4. Movements in every node of the real network.

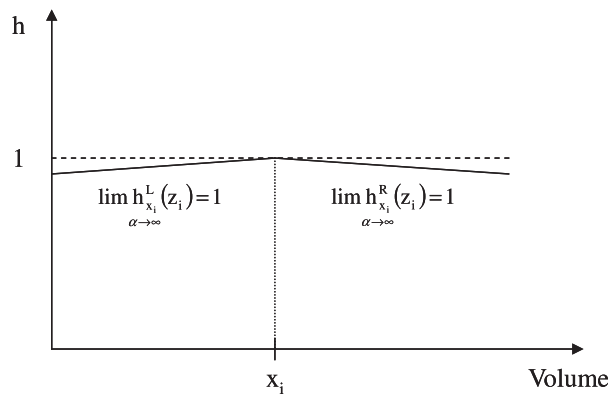


Fig. 5. Missing values' membership function.

omitted, and only two parameters have been used to verify the goodness of the adjustment: the lowest value of h and the sum of h_i .

The network has five intersections, of which three have six movements (intersections V, X and Z), while the other two have 12 potential movements (intersections W and Y). Data is available for all potential movements except for $v_1, v_2, v_3, v_4, y_6, y_7,$ and y_8 , whose values were lost. A special membership function with $h = 1.00$ always ($\alpha \rightarrow \infty$) was assigned to the lost values so that any adjusted value that met the boundary conditions would always have a membership grade of 1.00 (Fig. 5). Table 2 shows that for movements $v_1, v_2, v_3, v_4, y_6, y_7$ and y_8 , the value of h associated to the AV is always 1.00 for the three methods studied and for the hypothesis of equal or different α .

Columns 3, 5 and 7 in Table 2 show the AV and h_i using the three methods for $\alpha = 0.1$.

On the other hand, columns 4, 6 and 8 in Table 2 show the AV and h_i using the three methods for different α parameters depending on the reliability of the data. The α values used in this example were 0.2 for “W”, 0.3 for “Z”, and 0.1 for the rest.

As in the previous example, for any α parameter, the MM and the BO methods obtain the same and a higher value of minimum h ($h = 0.29$) than the MS method ($h = 0.00$). However, MS method obtains a higher value of $\sum h_i$ (40.14 versus 26.16 for

Table 2
Real intersection in the South of Spain: real base data with missing values, adjusted data, and results for different α ranges.

(1)	(2)	(3)		(4)		(5)		(6)		(7)		(8)	
		MS method	$\alpha = 0.1$	MS method	α^a	MS method	$\alpha = 0.1$	MS method	α^a	BO method	$\alpha = 0.1$	BO method	α^a
	OV	AV	h_i	AV	h_i	AV	h_i	AV	h_i	AV	h_i	AV	h_i
v_1^b	–	11,091	1.00	10,819	1.00	10,703	1.00	10,704	1.00	10,951	1.00	10,893	1.00
v_2^b	–	1764	1.00	1758	1.00	1743	1.00	1743	1.00	1497	1.00	1555	1.00
v_3^b	–	11,085	1.00	11,014	1.00	11,240	1.00	11,240	1.00	10,994	1.00	10,847	1.00
v_4^b	–	10,288	1.00	10,307	1.00	10,329	1.00	10,329	1.00	10,575	1.00	10,517	1.00
v_5	10,865	10,727	0.87	10,809	0.95	10,865	1.00	10,865	1.00	10,618	0.77	10,865	1.00
v_6	1207	1174	0.73	1201	0.95	1207	1.00	1207	1.00	1207	1.00	1207	1.00
w_1	5427	5669	0.56	5491	0.94	5427	1.00	5427	1.00	5445	0.97	5445	0.98
w_2	2714	2905	0.30	2719	0.99	2714	1.00	2714	1.00	2714	1.00	2714	1.00
w_3	3135	2914	0.30	3023	0.82	3135	1.00	3135	1.00	3135	1.00	3135	1.00
w_4	3123	3258	0.57	3179	0.91	3123	1.00	3124	1.00	3123	1.00	3123	1.00
w_5	3735	3764	0.92	3773	0.95	3735	1.00	3735	1.00	3735	1.00	3735	1.00
w_6	5601	5313	0.49	5311	0.74	5600	1.00	5600	1.00	5354	0.56	5207	0.65
w_7	695	744	0.30	695	1.00	695	1.00	695	1.00	695	1.00	695	1.00
w_8	3112	3182	0.78	3131	0.97	3112	1.00	3112	1.00	3112	1.00	3112	1.00
w_9	3880	3928	0.88	3907	0.97	3896	0.96	3896	0.98	3880	1.00	3880	1.00
w_{10}	505	470	0.31	502	0.97	505	1.00	505	1.00	505	1.00	505	1.00
w_{11}	310	289	0.32	307	0.95	310	1.00	310	1.00	310	1.00	310	1.00
w_{12}	904	841	0.30	913	0.97	904	1.00	904	1.00	904	1.00	904	1.00
x_1	4935	4588	0.30	4711	0.54	4812	0.75	4812	0.75	4800	0.73	4709	0.54
x_2	4725	5021	0.38	4788	0.87	4848	0.74	4848	0.74	4719	0.99	4715	0.98
x_3	7236	6739	0.31	6905	0.54	7236	1.00	7236	1.00	6990	0.66	6905	0.54
x_4	1809	1777	0.82	1736	0.60	1809	1.00	1809	1.00	1809	1.00	1747	0.66
x_5	4197	4394	0.53	4335	0.67	4197	1.00	4197	1.00	4344	0.65	4348	0.64
x_6	3350	3306	0.87	3197	0.54	3350	1.00	3350	1.00	3351	1.00	3197	0.54
y_1	1230	1176	0.56	1236	0.97	1230	1.00	1230	1.00	1230	1.00	1230	1.00
y_2	3700	3662	0.90	3717	0.95	3700	1.00	3700	1.00	3700	1.00	3700	1.00
y_3	4255	4555	0.29	4307	0.88	4257	1.00	4257	1.00	4555	0.29	4255	1.00
y_4	1410	1509	0.30	1441	0.78	1410	1.00	1410	1.00	1509	0.30	1410	1.00
y_5	2140	2076	0.70	2094	0.79	2140	1.00	2140	1.00	1990	0.30	2140	1.00
y_6^b	–	2320	1.00	2332	1.00	2369	1.00	2369	1.00	2369	1.00	2369	1.00
y_7^b	–	658	1.00	611	1.00	521	1.00	521	1.00	671	1.00	521	1.00
y_8^b	–	1527	1.00	1494	1.00	1691	1.00	1691	1.00	1679	1.00	1526	1.00
y_9	1150	1069	0.30	1131	0.83	1150	1.00	1150	1.00	1150	1.00	1150	1.00
y_{10}	310	289	0.32	324	0.65	310	1.00	310	1.00	310	1.00	310	1.00
y_{11}	1013	1044	0.69	1023	0.91	1013	1.00	1013	1.00	1013	1.00	1013	1.00
y_{12}	1410	1509	0.30	1442	0.77	1410	1.00	1410	1.00	1509	0.30	1410	1.00
z_1	4960	4975	0.97	4968	0.99	4962	1.00	4962	1.00	4962	1.00	4962	1.00
z_2	2120	2051	0.67	2104	0.97	2120	1.00	2120	1.00	2120	1.00	2120	1.00
z_3	1207	1122	0.30	1092	0.68	1207	1.00	1207	1.00	1122	0.30	1087	0.67
z_4	10,865	10,930	0.94	10,973	0.97	10,865	1.00	10,865	1.00	10,950	0.92	10,985	0.96
z_5	9660	9850	0.80	9906	0.92	9952	0.70	9952	0.90	9705	0.95	9952	0.90
z_6	6940	6451	0.30	6098	0.60	5870	0.00	5870	0.49	6451	0.30	5988	0.54
$g = \sum h_i$		26.16		36.50		40.14		40.85		35.97		38.61	
$h = \min h_i$		0.29		0.54		0.00		0.49		0.29		0.54	

Note: OV (observed value); AV (adjusted value).

^a $\alpha = 0.2$ for w_i ; $\alpha = 0.3$ for z_i ; $\alpha = 0.1$ for rest of cases.

^b $v_1, v_2, v_3, v_4, y_6, y_7$ and y_8 are missing values.

the MM method and 35.97 for the BO method). Thus, the results demonstrate that the BO method, while keeping the highest minimum of h , attains the best sum of h_i , so the best solution is chosen from among all the possibilities that satisfy the condition of maximizing the minimum h . Furthermore, introducing the analyst's knowledge of the different precisions of the data he is working with improves the results of the adjustment.

5. Summary and conclusions

The consistency of the observed traffic data is a concern because in nearly all cases traffic data contain some errors. The degree to which consistency must be satisfied depends on the purpose of the analysis. Processing observed data for consistency is crucial in an analysis where data interrelationships are important.

This paper proposes another step forward in using fuzzy logic optimization to obtain adjusted values. Two examples are given to present and explain the theoretical formulation and computational procedure. The proposed approach is robust enough to deal with other typical data discrepancies in transport situations. It preserves the integrity of observed data as much as possible, and allows the analyst to distinguish between reliable and less reliable data.

The approach is able to:

- Preserve the integrity of the observed data as much as possible. There are increasing concerns about data imputation and Base Data Integrity. The principle of Base Data Integrity is an important theme discussed by the American Society for Testing and Materials (ASTM, 1991) and the American Association of State Highway and Transportation Officials (AASHTO, 1992). The principle says that traffic measurements must be retained without modification and adjustment. Missing values should not be imputed in the base data. However, this does not prohibit imputing data at the analysis stage. In some cases, traffic counts with missing values could be the only data available for certain purposes and data imputation is necessary for further analysis. In accordance with the principle of Truth-in Data, AASHTO Guidelines (AASHTO, 1992) also recommends highway agencies to document the procedures for editing traffic data. For traffic counts with missing values, highway agencies usually either retake the counts or estimate the missing values. Estimating missing values is known as data imputation.
- Ensure flow consistency at any point in the network; the final estimate satisfies the law of flow conservation.
- Handle a large complicated network of any size and shape. The aim is to be able to solve any real problem, as shown in example 2.
- Handle data reliability; traffic-responsive control systems require reliable real-time information on the prevailing traffic counts to make sensible control decisions. This requisite is met by using the α parameter to define a different range for the membership function associated to each observed value.
- Limit the adjusted value within a tolerable deviation from the observed value, but allowing one tolerance for each value to be defined; this is achieved by using fuzzy logic and the definition of the α parameter.
- Be solved in a short computation time. The triangular membership function allows solving the problem using mixed linear programming.

The method is flexible so that it can handle cases in which data are questionable, some of the observed values are known and fixed ($\alpha = 0$), and there are considerable discrepancies in the observed data. The base of the membership function within which a feasible set of solutions is searched should be established according to the acceptable difference between adjusted and observed values.

Finally, the method is applicable to many other transportation problems in which consistency is important.

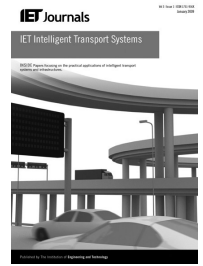
Acknowledgement

The authors appreciate the reviewers' comments and effort in order to improve the paper.

References

- AASHTO Guidelines for Traffic Data Programs, 1992. American Association of State Highway and Transportation Officials.
- American Society for Testing and Materials, ASTM, 1991. Standard Practice E1442, Highway Traffic Monitoring Standards, Philadelphia, PA.
- Chambers, R., 2001. Evaluation criteria for statistical editing and imputation. *National Statistics Methodological Series*, 28–41.
- Chakroborty, P., Kikuchi, S., 2003. Calibrating the membership functions of the fuzzy inference system: instantiated by car-following data. *Transportation Research Part C* 11 (2), 91–119.
- Geng, Y., Wu, X., 2008. The erroneous data imputation models for Beijing's urban traffic flow data by time series and correlation analysis. *Transportation Research Board 87th Annual Meeting*, Washington, DC, January 13–17, 2008 (on CD-ROM).
- Homaifar, A., McCormick, E., 1995. Simultaneous design of membership functions and rule sets for fuzzy controllers using genetic algorithms. *IEEE Transactions on Fuzzy Systems* 3 (1), 129–139.
- Kaczmarek, M., 2005. Fuzzy group model of traffic flow in street networks. *Transportation Research Part C* 13 (2), 93–105.
- Kikuchi, S., Miljkovic, D., 1999. Method to preprocess observed traffic data for consistency: application of fuzzy optimization concept. *Transportation Research Record* 1679, 73–80.
- Kikuchi, S., Miljkovic, D., Van Zuylen, H.J., 2000. Examination of methods that adjust observed traffic volumes on a network. *Transportation Research Record* 1717, 109–119.

- Kwon, J., Petty, K., Shieh, E., Kopelias, P., Papandreou, K., 2008. An automatic method for imputing and balancing link traffic counts. Transportation Research Board 87th Annual Meeting, Washington, DC, January 13–17, 2008 (on CD-ROM).
- Laaksonen, S., 1999. How to find the best imputation technique? Test with various methods. In: International Conference on Survey Nonresponse. Portland, Oregon, October 1999, pp. 28–31.
- Lee, S., Krammes, R.A., Yen, J., 1998. Fuzzy logic-based incident detection for signalized diamond interchanges. Transportation Research Part C 6 (5–6), 359–377.
- Marzano, V., Papola, A., Simonelli, F., 2008. Investigating the effectiveness of the o–d matrix correction procedure using traffic counts. Transportation Research Board 87th Annual Meeting, Washington, DC, January 13–17, 2008 (on CD-ROM).
- Pappis, C., Mamdani, W., 1977. A fuzzy logic controller for a traffic junction. IEEE Transactions on Systems Man and Cybernetics SMC-7, 707–717.
- Pentrice, G., 1987. Problems of present data collection and analysis. In: Proceedings Conference of the Institution of Civil Engineers, Institution of Civil Engineers, London.
- Rudy, K., Wang, H., Ni, D., 2008. Modeling and optimization of link traffic flow. Transportation Research Board 87th Annual Meeting, Washington, DC, January 13–17, 2008 (on CD-ROM).
- Sharma, S.C., Kilburn, P., Wu, Y.Q., 1996. The precision of AADT volumes estimates from seasonal traffic counts. Alberta example. Canadian Journal of Civil Engineering 23 (1), 302–304.
- Silva-Ramírez, E.L., 2007. Redes de Neuronas Artificiales en la edición e imputación de datos. Ph.D Thesis. University of Sevilla (in Spanish).
- Tussel, F., 2002. Neural networks and predictive matching for flexible imputation. In: DataClean 2002 Conference. Jyväskylä (Finland), pp. 29–31.
- Tzeu-Chen Han, 2008. Application of Fuzzy Regression on Air Cargo Volume Forecast. Transportation Research Board 87th Annual Meeting, Washington, DC, January 13–17, 2008 (on CD-ROM).
- Wang, L.-X., Mendel, J., 1992a. Back propagation of fuzzy systems as nonlinear dynamic system identifiers. In: Proceedings IEEE International Conference on Fuzzy Systems, San Diego, pp. 807–813.
- Wang, L.-X., Mendel, J., 1992b. Fuzzy basis functions. Universal approximation, and orthogonal least squares learning. IEEE Transactions on Neural Networks 3, 807–813.
- Wang, L.-X., Mendel, J., 1992c. Generating fuzzy rules by learning from examples. IEEE Transactions on Systems, Man and Cybernetics 22, 1414–1427.
- Zhong, M., Lingras, P., Sharma, S., 2004. Estimation of missing counts using factor, genetic, neural, and regression techniques. Transportation Research Part C 12 (2), 139–166.
- Zimmermann, H.J., 2001. Fuzzy Set Theory and Its Applications, fourth ed. Springer, Berlin.



Method to detect malfunctioning traffic count stations

J. de Oña¹ P. Gómez¹ E. Mérida-Casermeiro²

¹Transportation Engineering, Department of Civil Engineering, University of Granada, Granada, Spain

²School of Computer Science, Department of Applied Mathematics, University of Málaga, Málaga, Spain

E-mail: penelopegj@ugr.es

Abstract: This study presents a method for the automatic detection of malfunctioning traffic count stations (TCS) in a transport system. First, double linear optimisation is used to detect inadmissible errors in the recordings of a series of TCS and next, the TCS that are most likely to be failing are identified. The method has been applied to an urban traffic network showing success rates up to 93% in identifying the TCS that are failing.

1 Introduction

In traffic operation management and control field, accurate estimates of the density of vehicle flow density in road networks are very important. Information on traffic density may be ascertained from gross counts taken by loop detectors and other detection devices. However, the counts available may be incorrect owing to an improper collection process and errors.

When counting the number of vehicles that travel on a road, two types of errors can be committed:

1. *Admissible:* In general, admissible errors are the errors that are within the measuring device's tolerance and, therefore, they depend on the precision defined for each device by the manufacturer. For instance, if the manufacturer of the detectors in the traffic counts stations (TCS) indicates 3% reliability, it means that if one of the measurements is $x^{\text{obs}} = 784$, the real value $x^* \in [784(1 - 0.03), 784(1 + 0.03)]$. In practice, the admissible boundary of error tends to be somewhat higher, since margins tend to increase with use and over time.

2. *Inadmissible:* These are errors that not only give erroneous information, but also invalidate the work done. They can be due to detector malfunctioning (failure to record passing vehicles, constant recording of non-existent vehicles, always counting an arbitrary number etc.) or to failure on the part of the person who handles the detector (failure to set the counter to zero, erroneous readings etc.).

Let consider an intersection with two in (x_1 and x_2) and three out movements (x_3 , x_4 and x_5) the principle of flow conservation should verify that

$$x_1 + x_2 = x_3 + x_4 + x_5$$

Let the measurements be taken and the following is obtained:

- *Case 1:* $x_1^{\text{obs}} = 800$, $x_2^{\text{obs}} = 1200$, $x_3^{\text{obs}} = 600$, $x_4^{\text{obs}} = 700$ and $x_5^{\text{obs}} = 740$.

It is found that the above-mentioned condition is not verified, since $x_1^{\text{obs}} + x_2^{\text{obs}} = 2000$, whereas $x_3^{\text{obs}} + x_4^{\text{obs}} + x_5^{\text{obs}} = 2040$. Are the measurements reliable and therefore they can provide relevant information? Or are they indicating that a detector is failing and giving inadmissible measurements? In this case, and assuming that 3% of errors is admissible, we can indicate the existence of a set of values for the measurements that verifies the condition of conservation flow and is within the tolerance range: $x_1^{\text{adj}} = 808$, $x_2^{\text{adj}} = 1212$, $x_3^{\text{adj}} = 594$, $x_4^{\text{adj}} = 693$ and $x_5^{\text{adj}} = 733$. Therefore they should be close to the real values.

- *Case 2:* $x_1^{\text{obs}} = 800$, $x_2^{\text{obs}} = 1200$, $x_3^{\text{obs}} = 1600$, $x_4^{\text{obs}} = 700$ and $x_5^{\text{obs}} = 740$.

It is found that the above condition is not verified either, since $x_1 + x_2 = 2000$, whereas $x_3 + x_4 + x_5 = 3040$. However, at present no combination of x_i^{adj} values verifies flow conservation and falls within the 3% tolerance range. The inference would be that one of the measurements was erroneous and a detector must be repaired or replaced (unless there was a human error in the installation, reading or recording of the data).

A number of studies [1–5] attempt to find a solution to case 1 (admissible errors) to obtain adjusted data that are consistent with flow conservation laws.

For case 2 (inadmissible errors), several approaches [6–9] have been attempted to resolve or diminish count errors after they have been detected, but they do not address how they can be detected.

The methods for trying to detect errors may be classified according to the consistency criterion [10]:

1. *Fundamental consistency:* Data should be consistent with basic notions of traffic theory and should be physically plausible; establishes upper and lower boundaries for traffic

values (e.g. negative values and vehicle volumes that exceed the road's capacity cannot be measured).

2. *Network consistency*: Data should be related to measurements that are close in space and time. It is based on flow conservation when several connected nodes in a transport network are studied. This is the type of consistency shown in the preceding example.

3. *Historical consistency*: Historical observations can provide insight as to the plausibility of current data. Practice tells us that the values measured on a road are almost always given for an interval. Values outside of the interval may be plausible, but they indicate outliers, an anomaly that should alert the control service. The historical values constitute a basis for determining the boundaries of the interval in which normally consistent values must be found.

In current traffic control centres, detecting a malfunctioning count station is pseudo-automated because historical consistency marks the value interval each observation should have. If a measurement is not within that interval, an alarm is triggered, indicating a potential error in one of the TCS.

The problem arises when no historical values are available or when they exist but may indicate measurements as erroneous when they are actually correct. An incident on the network – repair work, accidents and weather issues, for instance – may alter track conditions significantly and cause outliers in the above-mentioned measurements without presupposing that the detector has failed, in fact there is a research field on this issue (among others [11–14]).

In Lin *et al.* [10] indicates several error detection techniques based solely on historical consistency. They do not take nearby detectors, that is, network consistency, into consideration. Other approach is to incorporate observations from adjacent detectors [4]. This paper presents a method that is complementary to the existing ones, where basic consistency and network consistency are taken into consideration.

The method automatically detects a TCS that is failing, by only considering the data observed by the network detectors as input data. Once the detector that is failing has been detected, the procedure can be repeated to see if the remaining measurements are consistent and free of errors.

This paper is organised as follows: Section 2 describes the method and the computational issues; in Section 3 the method is applied to an urban network; Section 4 discusses the effect of the model's variables on the results; and finally, Section 5 presents the main conclusions of the paper.

2 Methodology

The method presented in this paper to detect and identify a malfunctioning detector is based on the resolution of a linear programming (LP) problem. In general terms, the \mathbb{R}^n region that meets certain restrictions is known as the LP's feasible region. That is what will be built for the problem posed in this paper.

2.1 Feasible region

Let a series of measurements be taken $\{x_i^{\text{obs}}\}$ and that the tolerance indicated for each measurement is α_i . This tolerance is usually expressed as a percentage of the measured value, since it is reasonable to assume that any absolute errors incurred will be lower for small magnitudes

than for larger ones, assuming the detectors function under the conditions specified by the manufacturer: $\forall i; x_i^* \in [a_i, b_i]$, where $a_i = x_i^{\text{obs}} - \alpha_i x_i^{\text{obs}}$ and $b_i = x_i^{\text{obs}} + \alpha_i x_i^{\text{obs}}$. In Example 1, a 3% error was considered admissible for all the measurements, and therefore we would take $\forall i, \alpha_i = 3\%$, although in other cases a different error for each detector could be considered.

Given a set of observed values $\{x_i^{\text{obs}}\}, i \in \mathcal{I}$, (where \mathcal{I} is a set of indexes) each with a tolerance of α_i , we define the 'admissible region' as the set $\mathcal{A} \subset \mathbb{R}^n$, such that $\forall \vec{x} = \{x_i\} \in \mathcal{A}$ where the following conditions are satisfied:

1. $x_i^{\text{obs}} - \alpha_i x_i^{\text{obs}} \leq x_i \leq x_i^{\text{obs}} + \alpha_i x_i^{\text{obs}}$.
2. Vector \vec{x} verifies flow conservation laws.

Attention should be paid to the fact that the cardinal of the set of observed values and the number of variables, n , do not necessarily coincide. Thus, to continue with the previous example, let the set of observed values be $x_1^{\text{obs}} = 800, x_2^{\text{obs}} = 1200, x_3^{\text{obs}} = 600$ and $x_4^{\text{obs}} = 700$, which would give the admissible region

$$\mathcal{A} = \{\vec{x} \in \mathbb{R}^5 / 776 \leq x_1 \leq 824, 1164 \leq x_2 \leq 1236, 582 \leq x_3 \leq 618, 679 \leq x_4 \leq 721, x_5 = x_1 + x_2 - x_3 - x_4\}$$

where the first four intervals are obtained by $x_i = x_i^{\text{obs}} \pm \alpha_i x_i^{\text{obs}} = x_i^{\text{obs}}(1 \pm \alpha_i)$, adding the flow conservation law: $x_1 + x_2 = x_3 + x_4 + x_5$.

Theorem 1: If all the detectors function properly, the feasible region is not empty ($\mathcal{A} \neq \emptyset$).

Obviously, if all the detectors give admissible errors, then the true values vector, \vec{x}^* , belongs to the feasible region ($\vec{x}^* \in \mathcal{A}$).

Therefore the inference is

Corollary 1: Si $\mathcal{A} = \emptyset$, one of the detectors is giving an inadmissible error.

Corollary 1 provides a method for detecting incorrect measurements by taking into consideration fundamental inconsistencies and network inconsistencies. Although the converse theorem is not true, that is,

A detector may produce an inadmissible error, but the remaining detectors' margins permit admissible values and, therefore, $\mathcal{A} \neq \emptyset$. In practice, this means that although there exists out of range measures, it is possible even to obtain a consistent vector. Hence, if a detector is severely malfunctioning it will be impossible to generate consistent traffic counts.

We should also consider that if there are several vectors in \mathcal{A} , ($\mathcal{A} \neq \emptyset$), some are more plausible than others, insofar as they are closer to the observed values. Hence, for a vector $\vec{x} \in \mathcal{A}$ we can associate another vector $\vec{h} = \{h_i\}$ such that the verisimilitude of the i th component is

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

Fig. 1 shows the verisimilitude of assigning a value x_i when x_i^{obs} with reliability α_i has been observed.

For the sake of simplicity, a triangular shape function has been chosen since the function shape is not an important issue, since the aim is to check if the adjusted value is in or out of the feasible region and simplicity of linear decay

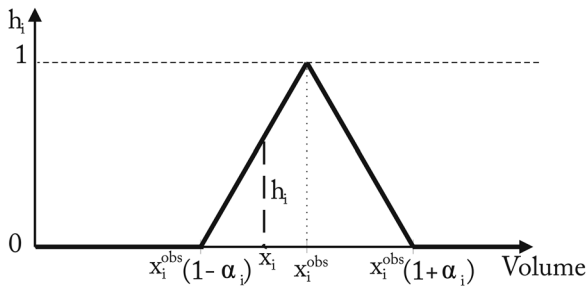


Fig. 1 Verisimilitude function for a single observation

allows it to be solved by linear programming. However, other polygonal function could be used, as it is stated in [1].

Assuming $x_i^{obs} > 0, \forall i \in \mathcal{I}$ and making the relevant transformations in (1), finding out whether an admissible set of values exists becomes a problem of finding out whether a solution to the linear optimisation problem exists:

Problem 1

$$\text{Maximise: } \sum_{i \in \mathcal{I}} h_i$$

$$\text{Subject to: } \begin{cases} 0 \leq h_i \leq 1, & x_i \geq 0 \\ x_i + \alpha_i x_i h_i \leq x_i^{obs}(\alpha_i + 1) \\ -x_i + \alpha_i x_i h_i \leq x_i^{obs}(\alpha_i - 1) \\ \mathbf{M}\vec{x} = 0 \end{cases}$$

where x_i, h_i and h are the variables that can be considered adjusted (consistent) values, variable verisimilitude and minimal verisimilitude, respectively, and where the flow conservation laws are represented by the homogeneous linear system $\mathbf{M}\vec{x} = 0$. Thus, for case 1 with the single conservation law: $x_1 + x_2 - x_3 - x_4 - x_5 = 0$, the matrix $\mathbf{M} = (1, 1, -1, -1, -1)$. In general, the matrix \mathbf{M} will have as many rows as existing flow conservation equations. Very different target functions could have been selected for this task, but this will also serve the second aim of this paper: To determine which detector is producing erroneous values. The benefit of transforming the problem into a linear programming problem is being able to count on multiple and optimised routines for the solution. (See <http://www.mathworks.com/help/toolbox/optim/ug/linprog.html> [15].) It is easy to amend the above method by considering different margins to the right and to the left of the observed values, that is, $x_i^{obs} \in (x_i^{obs} - \alpha_i^L x_i^{obs}, x_i^{obs} + \alpha_i^R x_i^{obs})$.

2.2 Detection of inadmissible measurements

Let the problem of resolving linear programming 1 in Section 2.1 be posed and that there is no solution, since $\mathcal{A} = \emptyset$. We would be in the case of Corollary 1, which indicates that one of the measurements is inadmissible. Unfeasible should not be confused with outliers, since the latter may be correct and owing to traffic anomalies (an accident, repairs etc.) but consistent with flow conservation laws. To detect an incorrect measurement, we relax the manufacturer's α_i margins, multiplying them by a constant $K \gg 0$ so the new linear optimisation problem will have a non-empty admissible region. That is,

Problem 2

$$\text{Maximise: } \sum_{i \in \mathcal{I}} h_i$$

$$\text{Subject to: } \begin{cases} 0 \leq h_i \leq 1, & x_i \geq 0 \\ x_i + K\alpha_i x_i h_i \leq x_i^{obs}(K\alpha_i + 1) \\ -x_i + K\alpha_i x_i h_i \leq x_i^{obs}(K\alpha_i - 1) \\ \mathbf{M}\vec{x} = 0 \end{cases}$$

It is known that one property of the 'maxsum' objective function is that it gives high values to most variables at the expense of giving low values to a few variables [16]. In this case, its effect is to assign values very close to the observed ones (high verisimilitude) to the detriment of assigning very distant values to a few (low verisimilitude). The measurement that produces $h = \min\{h_i\}$ in problem 2 will be proposed as inadmissible. We can always obtain a K that is large enough to make $\mathcal{A} \neq \emptyset$, since its effect is to increase the variables' admissible margin. In an extreme case, any measurement x_i would fit into the $(x_i^{obs} \pm K\alpha_i x_i^{obs})$ interval. It could be assumed that selecting K would modify the solution obtained, but the following theorem shows that such is not the case.

Theorem 2: If the problem 2 is solved by using two different values for K ($K_1 \neq K_2$), performing both feasible solutions, then optimum solutions for K_1 and K_2 verify:

1. The optimum vector $\vec{x}^{(1)*}$ for K_1 is also optimum vector for K_2 : $\vec{x}^{(2)*} = \vec{x}^{(1)*}$
2. The index of observation with minimum value for h_i is the same for both constants: $\arg \min_i \{h_i^{(1)}\} = \arg \min_i \{h_i^{(2)}\}$

Proof is given in Appendix.

2.3 Proposed algorithm

From previous considerations, Fig. 2 is proposed.

Fig. 2 is focused on detecting inadmissible observations from the network consistency viewpoint. However, it is easy to incorporate any available additional information. For instance, by changing the upper bound of any variable (adding the restriction $x_i \leq U_i$ in step 3b), or by changing the lower bound of any variable, which by default is 0 ($x_i \geq L_i$) etc. This fact allows making it suitable to perform fundamental consistency, generally expressed by bounds.

This method could be complementary to standard pre-process that analyses historical consistency [10]. That is, observed variables must be into a real interval, in other case the observation is considered an outliers. An outliers must be analysed separately since it can be produced by anomalous traffic, but be correct.

Perhaps Fig. 2 was only executed to verify that the detectors were working properly, but it is usually part of the study on a region's traffic. In such case, the next step would be to obtain the adjusted data, that is, the consistent data that most closely resembles the observed data. Any data deemed inadmissible during the pre-process will have been eliminated from the observed data using one of the procedures suggested by other authors [1-4].

3 Application to an urban network

3.1 Road network data

The method is applied to the urban network shown in Fig. 3.

Algorithm 1

- 1) Read values for $\alpha_i^L, \alpha_i^R, y, x_i^{obs}$
- 2) Represent the flow conservation laws by matrix M .
- 3) Repeat until all $h_i^* > 0$:
 - a) Represent all inequalities by matrix A and vector \vec{b} :

$$x_i + \alpha_i^R x_i h_i^* \leq x_i^{obs} (\alpha_i^R + 1)$$

$$-x_i + \alpha_i^L x_i h_i^* \leq x_i^{obs} (\alpha_i^L - 1)$$
 - b) Express restrictions $x_i \geq 0$.
 - c) Solve LP with the target function Maximise $\sum_i h_i^*$.
 - d) If all $h_i^* \geq 0$, go to step 4, else:
 - d1) Evaluate $h^* = \min_i h_i^*, K = \frac{1-h^*}{0.9}$
 - d2) Replace $\alpha_i^R \leftarrow K \alpha_i^R$ and $\alpha_i^L \leftarrow K \alpha_i^L$, into A and \vec{b} (step 2a).
 - d3) Solve LP with the objective function Maximise $\sum_i h_i^*$.
 - d4) The index k that produces $h_k^* = \min_i h_i^*$ is obtained.
 - d5) Observation x_k^{obs} is ellipsed and considered as erroneous.
 - d6) Return to step 3). With initial values of α_i^R and α_i^L , but the ellipsed one.
 - e) Finish. (Ellipsed observations are considered as inadmissible ones.)

Fig. 2 Erroneous sensor detector

The network has seven intersections, of which four have 12 movements (intersections D, E, F and G), two have six movements (A and C), whereas the last one has five potential movements (intersection B). Hence, in total, there are 86 unknown variables. As it is impossible to guarantee that a set of true data will always be available, the initial set of data will be a set of consistent data that is very close to the observed data.

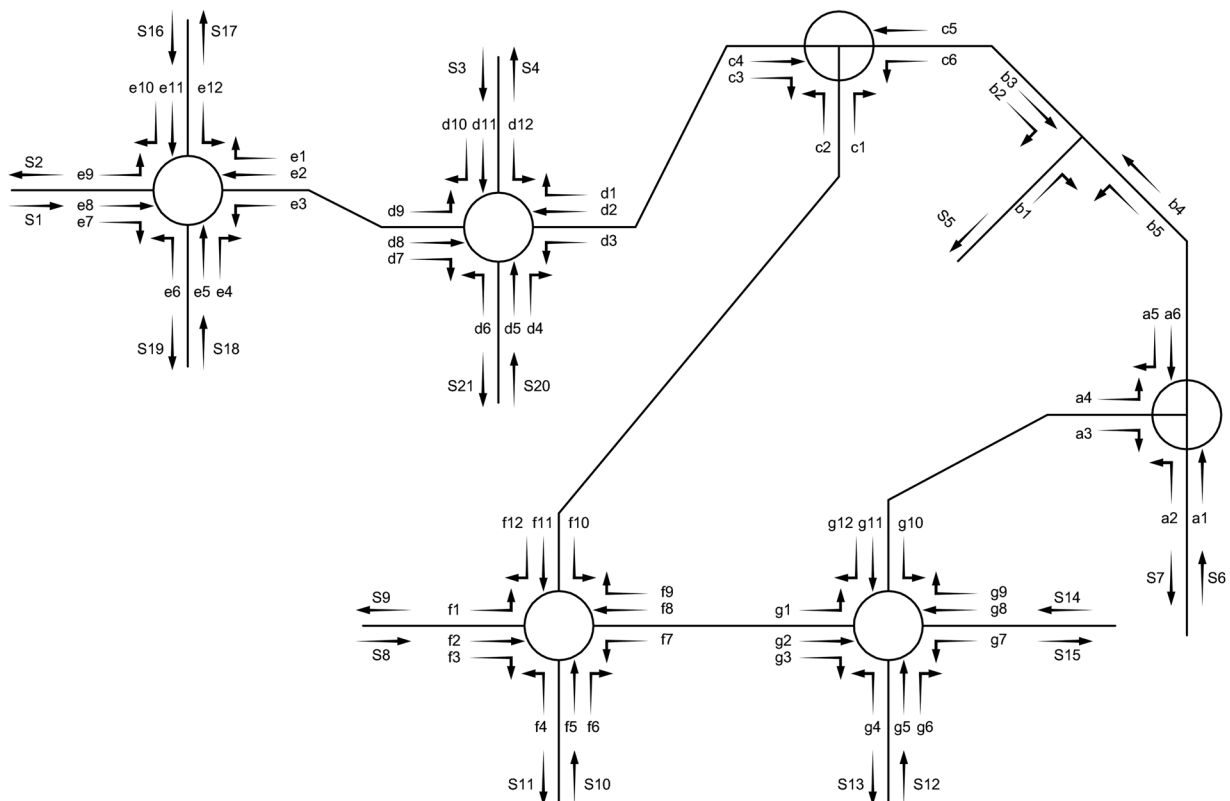


Fig. 3 Example of an urban network

Consider the situation shown in Fig. 3, in which consistent true data are available (theoretical values – TVs), where the data that comply with flow conservation in the traffic network concerned are deemed to be consistent. In other words, the sum of incoming vehicles is equal to the sum of outgoing vehicles at any network intersection.

This consistent database is used to randomly deform values, by a uniform distribution, with a tolerance of $\pm 3\%$, which is the tolerance shown by the count stations most commonly used in urban networks [17]. This is not deterministic, however, because if the detector was of another type or had a different tolerance, a value other than $\pm 3\%$ could be considered. The model allows a different α for each observed datum to be defined (several types of detectors with different tolerances). As shown in Section 2, it even permits the definition of asymmetric feasible regions.

Having obtained a randomly distorted database within the above-mentioned tolerance, it could then be considered as the data that would be obtained in an ideal counting campaign in which all 86 potential movements would be measured. Therefore it could be taken as the series of observed data in an urban network (observed values – OVs). In this case, the values would not be consistent, according to the above definition (the sum of incoming vehicles would not be equal to the sum of outgoing vehicles).

The fact that a base of consistent true data is used and subsequently randomly distorted permits a comparison between the results obtained and real life, and verification of the goodness of the method proposed.

3.2 Results

OV obtained randomly from TV with a tolerance of $\pm 3\%$ is used to verify the goodness of the model. Next, a datum is randomly selected and distorted to simulate a detector error

Table 1 Obtained results with a simulated error of 75, 50, 25 and 10%

Percentage simulated error	Error is detected		Error is pointed out and gets by h_{\min}		Error is pointed out and gets by second h_{\min}		Error is pointed out in total
	Number of times (A)	Proportion A/500	Number of times (B)	Proportion B/500	Number of times (C)	Proportion C/500	Success proportion (B + C)/500
75	491	0.982	443	0.886	23	0.046	0.932
50	486	0.972	386	0.772	37	0.074	0.846
25	438	0.876	269	0.538	27	0.054	0.592
10	266	0.532	137	0.274	24	0.048	0.322

that exceeds the error specified by the manufacturer or, in other words, a deviation from the detector's allowed tolerance. Specifically, deviations of 75, 50, 25 and 10% from OV are simulated.

This deformation gives an initial database for each example generated (each of that contains an erroneous datum). For each one of the four deviations, 500 examples are randomly generated from OV. In all, 2000 examples are executed. Table 1 shows the results for the random examples.

Column 1 in Table 1 shows the simulated error in a randomly selected measurement apparatus. Columns 2 and 3 show the number of times an error is detected in all the random samples. That is, the number of times $\mathcal{A} = \emptyset$ is obtained applying Theorem 1. Column 2 points out the number of times $\mathcal{A} = \emptyset$ is obtained for the random examples, which is when the adjusted value lies outside of the detector's allowed tolerance, and outside of the set boundaries of the feasible region. This indicates that a detector is giving a value that is higher than the allowed deviation, which in turn means that a detector is failing. Column 3 shows the same thing in relative terms.

By increasing all α_i from 0.03 [(see step d1) in Fig. 2] in a two steps process, the feasible region is extended in order to allow $\mathcal{A} = \emptyset$ to be found for every i . This value was selected because it produces all $h_i^* > 0$ at next step, as can be deduced from (2) in proof of Theorem 2.

Table 1, column 4 shows the number of times the index i that produces $h = \min h_i$, coincides with the failing TCS. Columns 6 and 7 show the number of times (and proportion, respectively) in which the failing TCS is the one that shows the second lowest value. Hence, when a TCS perform a 75% of error, it coincides with the error obtained by the second minor value of h_i in 5% of cases.

Column 8 shows the proportion of times that the model is able to detect the failing detector (i.e. adding the number of times it detects the detector that fails, whether it is the h_i minimum or the value immediately above it). This result points out the proportion of times at which it indicates a detector that is failing, out of all the random examples. This is the model's proportion of success, and it is calculated by adding column 4 and column 6, and dividing by the total number of examples simulated. For an error of 75%, the success rate is 93%. For the remaining cases, the model finds that there is a malfunctioning detector, but it does not point it out in the first or second places.

Table 1 shows that the model's success increases in the same measure as the device's error increases and worsens as the error diminishes, and the closer it is to the measurement device's tolerance range.

If the ratio (r) is expressed as the proportion of times that an error is detected compared to the number of examples

executed (Table 1, column 2), the ratio of cases in which a failing detector is detected for each simulated error can be compared.

In other words, if N random examples have been executed (in this case, $N = 500$) and A times errors have been detected (Table 1, column 2), the estimated ratio obtained experimentally is $r = A/N$ (Table 1, column 3). In this manner, for an error of 75%, the error is detected in 98% of cases; for an error of 50%, in 97% of cases; for an error of 25%, in 88% of cases; and finally, for a simulated error of 10%, an error is detected in 53% of cases.

4 Sensibility analysis to different variables

First, the effect of the situation of the failing traffic counts will be analysed. Second, what happens when certain points of the network have not been counted? Finally, the sensitivity to the number of not counted data in the network will be analysed (with $\sim 50\%$ more and 50% less points not counted).

4.1 Effect of the situation of the failing detector

How does the sensitivity depend on which detector is malfunctioning? In order to analyse the method's sensitivity to the detector position, a selective choosing of the malfunctioning detector has been made. At first stage, for each scenario, the model was forced to choose an edge detector (S_1, S_2, \dots, S_{21} , or b_1 in Fig. 3), and at second stage the central ones (the remaining detectors) have been chosen to be failing. Table 2 shows the results.

The method detects an error on the edge of the network better than when the detector is situated in the centre. This is logical because of the following reason: when an edge detector is getting an inadmissible error, whereas the rest adjacent measurements are corrects, must significantly modify its value in order to reach network consistency. That is because a small amount of adjacent detectors exists which can be modified within the margin established by the feasible region. On the other hand, a major modification of these adjacent detectors makes the constraints able to be affected; therefore the $\sum_i h_i$ is reduced. The target function forces to modify the one that is giving an erroneous measurement.

Whereas, for centre detectors, the measurements are linked to more variables that can be modified within the feasible region. Hence, an inadmissible small error (around 10%) is easier to count on the adjacent values margin and move all of them, in order to obtain all measures within its feasible region, than a big change in the malfunctioning detector.

Table 2 Results for centre and edge detectors with a simulated error of 75, 50, 25 and 10%

Percentage simulated error	Error is detected		Error is pointed out and gets by h_{min}		Error is pointed out and gets by second h_{min}	
	Centre	Edge	Centre	Edge	Centre	Edge
	75	492	500	426	500	24
50	474	500	340	486	49	6
25	419	500	227	428	32	25
10	205	466	90	249	11	45

4.2 Effect of points that are not counted

In this subsection the effect of movements that have not been counted is analysed.

Presumably, the network in Fig. 3 shows seven movements that have not been counted (movements $c_2, c_3, c_4, d_9, d_{10}, d_{11}$ and d_{12}). This implies around 8% of all the movements in the network. This percentage is considered to be normal in counting campaigns in a traffic network [18]. A case consisting of 500 random examples is simulated below, in which the number of not measured movements is increased 50% (10 not measured movements), followed by a case in which the number of not measured points is diminished in 50% (four not measured movements).

Table 3 and Fig. 4 show a comparison between the results obtained in the study with four hypotheses (all measured data, 4, 7 and 10 not measured data). In Fig. 4 the x -axis represents the simulated distortions for the measurement device and the y -axis represents the proportion of times the error is detected.

Taking column 3 ($A/500$) in Tables 1 and 3 into consideration, a comparison can be made about the number of times an error is detected in each case. Table 3 shows that the ratio of errors detected for the simulated scenarios gradually diminish when there is less measured data available (i.e. less information).

From Fig. 4, Tables 1 and 3, it is possible to analyse the model’s sensitivity to the number of not measured movements in a case where all the data from all the TCS (i.e. all measured data) is available. The x -axis represents the simulated percentage of the device error (10, 25, 50 and

75%) and the y -axis data show the percentage of success for every case, in comparison with the one in which all the data are measured. In the event that a 75% error occurs in a detector, for instance, the chart will show that the model presented in this paper is 93% successful if 4 network data are not measured, 90% if 7 network data are not measured and 84% if 10 data are not measured.

Thus, the conclusion would be that the model gives good success results even when the number of not measured data increases, although, obviously, when more data are available, the more it improves.

4.3 Combined effect of the size of the error and the number of points that are not counted

Fig. 4 shows the ascending trend of the ratio when a detector’s error increases in all the hypotheses. The trend is even more pronounced when it moves from an error close to the detector’s tolerance range (such as 10%) to around 25%, after which the detector’s behaviour is asymptotic, reaching an error ratio within the range 0.9–1 for the biggest device error simulated. In other words, when the error exceeds the threshold at around 25%, it can be asserted that the model succeeds in around 90% of the cases.

In Fig. 4 the $1 - \sigma$ errors bars have been included in order to show conclusions do not owe to random.

Table 4 showed the ratios (or proportion of success, p_i) at which error is detected in every scenario. To demonstrate that the model’s proportion of success increases when more data are measured ($p_{i+1} < p_i$) and that the observed results

Table 3 Results with a simulated error of 75, 50, 25 and 10% with 4, 7 and 10 not measured movements

Percentage simulated error	Error is detected		Error is pointed out and gets by h_{min}		Error is pointed out and gets by second h_{min}		Error is pointed out in total
	Number of times (A)	Proportion $A/500$	Number of times (B)	Proportion $B/500$	Number of times (C)	Proportion $C/500$	Success proportion $(B + C)/500$
	<i>4 not measured movements</i>						
75	466	0.932	397	0.794	27	0.054	0.848
50	447	0.894	353	0.706	36	0.072	0.778
25	391	0.782	234	0.468	38	0.076	0.544
10	242	0.484	95	0.190	29	0.058	0.248
<i>7 not measured movements</i>							
75	449	0.898	374	0.748	30	0.060	0.808
50	425	0.850	315	0.630	42	0.084	0.714
25	351	0.702	201	0.402	28	0.056	0.458
10	226	0.452	100	0.200	30	0.060	0.260
<i>10 not measured movements</i>							
75	418	0.836	360	0.720	22	0.044	0.764
50	372	0.744	282	0.564	35	0.070	0.634
25	338	0.676	211	0.422	29	0.058	0.480
10	203	0.406	87	0.174	28	0.056	0.230

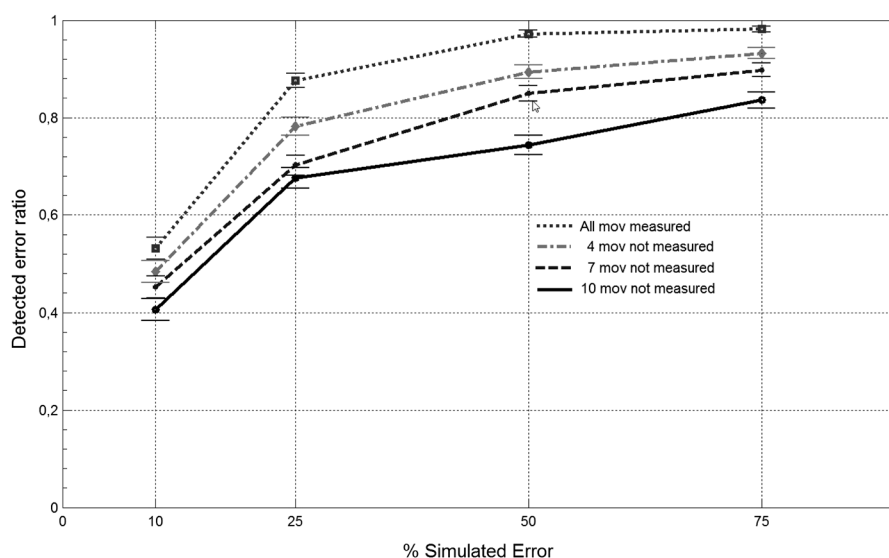


Fig. 4 Sensitivity analysis of success against increase in the number of data not measured

Table 4 Ratios calculated for every scenario providing the standard deviation

Percentage of simulated error	All movements are measured		4 not measured movements		7 not measured movements		10 not measured movements	
	Proportion	σ	Proportion	σ	Proportion	σ	Proportion	σ
75	0.982	0.006	0.932	0.011	0.898	0.014	0.836	0.017
50	0.972	0.007	0.894	0.014	0.850	0.016	0.744	0.020
25	0.876	0.015	0.782	0.018	0.702	0.020	0.676	0.021
10	0.532	0.022	0.484	0.022	0.452	0.022	0.406	0.022

Table 5 Test of hypotheses

% error	p_i	P_{i+1}	Z_{exp}
<i>all measured against 4 not measured movements</i>			
75	0.982	0.932	3.927110655
50	0.972	0.894	4.993847666
25	0.876	0.782	3.978624361
10	0.532	0.484	1.519839919
<i>4 against 7 not measured movements</i>			
75	0.932	0.898	1.931244679
50	0.894	0.850	2.086907096
25	0.782	0.702	2.903158213
10	0.484	0.452	1.014529379
<i>7 against 10 not measured movements</i>			
75	0.898	0.836	2.898969254
50	0.850	0.744	4.20341134
25	0.702	0.676	0.888433399
10	0.452	0.406	1.471128409

Significant cases in bold

are not due to chance, a hypotheses of proportional difference was tested at a significance level of 5%, taking $N_{i+1} = N_i = 500$ [19, 20].

Three statistical tests were conducted to compare the three hypotheses in groups of two. That is, first hypothesis of all movement measured was tested against 4 not measured movements, the case of 4 not measured movements against 7 not measured data, and last, 7 not measured data were tested against 10 not measured data. The $Z_{exp} = (p_i - p_{i+1})/\sigma$ is calculated and compared with the $Z_{theoretical} = 1.645$, it

determines the significant region ($Z_{exp} > 1.645$). The results are given in Table 5.

It is found that $p_{i+1} < p_i$ in all cases and statistically significant results are obtained for the cases of 75, 50 and 25% error in the first and second tests, and in the third one the results are significant after the 50% error.

Therefore it can be asserted that the success proportion improves with the number of counted data and this fact is not due to chance.

5 Summary and conclusions

This paper presents a method for detecting inadmissible errors in TCS and identifying which device is more likely to be failing. The method is based on a double linear optimisation process that can easily be solved with existing software on the market, and which we consider highly useful for practitioners.

If the method detects the existence of an inadmissible error in the TCS' measurements when the first linear optimisation is used, a second optimisation can be used so the method can obtain the detector that is most likely to be failing (the one that obtains the $\min_i h_i$). This facilitates to replace or fix them for obtaining adjusted data.

Four different cases of potential errors were simulated in order to identify the effects on the method (deviations of 10, 25, 50 and 75%). The results show that the method works better with bigger errors (75%), which are more frequent when dealing with malfunctioning detectors, than with small errors (10%), close to the TCS's tolerance (3%). For deviations of around 25% of their theoretical value, the

method is 88% efficient for detecting that there is an error in the measures. The efficiency in identifying a failing detector can be considered good (over 90%) when the error is over 75% of the deviation, and diminishes as the errors become smaller.

The same tolerance was considered for all the TCS (3%), but the model is versatile and allows assigning a different tolerance to each detector according to its type and level of precision.

Finally, a statistical test has been conducted to demonstrate that the increase in the number of times an error is detected when more movement counts were obtained as opposed to a gradually decreasing number of times is not due to chance. This serves to assert that the results are significant and the size of the sample selected is sufficient to corroborate the conclusions arrived at in this paper.

Usually studies perform automated data checking by comparing measured data to historical data for consistency [10]. Sometimes, however, there are no historical data and only the observed database is available. This is when the method proposed in this paper becomes a good tool for detecting errors, since the only incoming data required are the observed data, with no need for preprocessing. Actually, both approaches could be considered as complementary: it is possible to use fundamental and network consistency for detecting inadmissible errors and, historical consistency as alarm signal.

6 Acknowledgment

The authors appreciate the reviewers' comments and effort in order to improve the paper.

7 References

- 1 Kikuchi, S., Miljkovic, D.: 'Method to preprocess observed traffic data for consistency: application of fuzzy optimization concept', *Transp. Res. Rec.*, 1999, **1679**, pp. 73–80
- 2 Wall, Z.R., Dailey, D.J.: 'Algorithm for detecting and correcting errors in archived traffic data', *Transp. Res. Rec.*, 2003, **1855**, pp. 183–190
- 3 Vanajahshi, L., Rillet, L.R.: 'Loop detector data diagnostics based on conservation-of vehicles principle', *Transp. Res. Rec.*, 2004, **1870**, pp. 162–169
- 4 de Oña, J., Gómez, P., Mérida-Casermeyro, E.: 'Bilevel fuzzy optimization to pre-process traffic data to satisfy the law of flow conservation', *Transp. Res. Rec. C*, **19**, (1), pp. 29–39
- 5 Schleifer, W., Mannle, M.: 'Online error detection through observation of traffic self-similarity', *IEE Proc., Commun.*, 2001, **148**, (1), pp. 38–42
- 6 Nihan, N.L., Davis, G.A.: 'Application of prediction-error minimization and maximum likelihood to estimate intersection O-D matrices from traffic counts', *Transp. Sci.*, 1987, **23**, pp. 77–90
- 7 Nihan, N.L., Davis, G.A.: 'Recursive estimation of origin-destination matrices from input/output counts', *Transp. Res. B*, 1987, **21**, pp. 149–163
- 8 Nihan, N.L., Davis, G.A.: 'Application of prediction-error minimization and maximum likelihood to estimate intersection O-D matrices from traffic counts', *Transp. Sci.*, 1989, **23**, pp. 77–90
- 9 Tavana, H., Mahmassani, H.: 'Estimation of dynamic origin-destination flows from sensor data using bi-level optimization method'. Proc. 80th Annual Meeting of the Transportation Research Board, CD ROM, 2000
- 10 Lin, D.-Y., Boyles, S., Valsaraj, V., Waller, S.T.: 'Fuzzy reliability assessment for traffic data', *J. Chin. Inst. Eng.*, 2012, **35**, (3), pp. 1–14

- 11 Thomas, T., van Berkum, E.C.: 'Detection of incidents and events in urban networks', *IET Intell. Transp. Syst.*, 2009, **3**, (2), pp. 198–205
- 12 Zhang, H.-Z., Wang, J., Zi-hui Ren, Z.-H.: 'Rough sets and FCM-based neuro-fuzzy inference system for traffic incident detection'. ICNC'08. Fourth Int. Conf. on Natural Computation, 2008, vol. 7, pp. 260–264
- 13 Srinivasan, D., Sanyal, S., Sharma, V.: 'Freeway incident detection using hybrid fuzzy neural network', *IET Intell. Transp. Syst.*, 2007, **1**, (4), pp. 249–259
- 14 Tang, S., Gao, H.: 'Traffic-incident detection-algorithm based on nonparametric regression', *IEEE Trans. Intell. Transp. Syst.*, 2005, **6**, (1), pp. 38–42
- 15 LINPROG. Available at <http://www.mathworks.com/help/toolbox/optim/ug/linprog.html>. Last accessed: 22 March 2011
- 16 Saameño Rodríguez, J.J., Guerrero Garca, C., Muñoz Pérez, J., Mérida Casermeyro, E.: 'A general model for the undesirable single facility location problem', *Oper. Res. Lett.*, 2006, **34**, (4), pp. 427–436
- 17 CEGASA. Available at <http://www.cegasatrafic.com/es/listado/b2/TomaDatos.html>. Last accessed: 30 September 2010
- 18 Zhong, M., Lingras, P., Sharma, S.: 'Estimation of missing traffic counts using factor, genetic neural and regression techniques', *Transp. Res. C*, 2004, **12**, pp. 139–166
- 19 Anderson, T.W., Sclove, S.T.: 'The statistical analysis of data' (The Scientific Press, 1986, 2nd edn.)
- 20 Mendenhall, W., Sincich, T.: 'Statistics for the engineering and computer sciences' (Dellen Publishing Company, San Francisco, CA, USA, 1988, 2nd edn.)

8 Appendix: Proof of Theorem 2

Let $\vec{x} \in \mathcal{A}$ and $\forall i, h_i^{(1)} = 1 - (|x_i - x_i^{\text{obs}}|/K_1 \alpha_i x_i^{\text{obs}})$, $h_i^{(2)} = 1 - (|x_i - x_i^{\text{obs}}|/K_2 \alpha_i x_i^{\text{obs}})$ then it is easily obtained

$$K_1(h_i^{(1)} - 1) = K_2(h_i^{(2)} - 1) \quad (2)$$

and by naming m the number of observed variables

$$\begin{aligned} \sum_{i \in \mathcal{I}} \{h_i^{(1)}\} - m &= \sum_{i \in \mathcal{I}} \{h_i^{(1)} - 1\} = \frac{K_2}{K_1} \sum_{i \in \mathcal{I}} \{h_i^{(2)} - 1\} \\ &= \frac{K_2}{K_1} \left(\sum_{i \in \mathcal{I}} \{h_i^{(2)}\} - m \right) \end{aligned} \quad (3)$$

From (3), it can be defined the monotonically increasing function: $S_1 = (K_2/K_1)(S_2 - m) + m$ between $S_1 = \sum_{i \in \mathcal{I}} h_i^{(1)}$ and $S_2 = \sum_{i \in \mathcal{I}} h_i^{(2)}$.

If we assume that $\vec{x}^{(1)*} \in \mathcal{A}$ is the set of values that produces the optimal solution to problem 2 with K_1 , producing values for target functions S_1^* and S_2^* for the constant K_2 . Then, if a $\vec{x}' \neq \vec{x}^{(1)*} \in \mathcal{A}$ exists and could provide a better solution to problem 2 with K_2 ($S_2' > S_2^*$), then the monotony of the equation produces that for this vector \vec{x}' , ($S_1' > S_1^*$), which is absurd, since no solution can be better than the optimal solution. Therefore there cannot exist any vector \vec{x}' giving a better value to the target function of problem 2 with K_2 than $\vec{x}^{(1)*}$. This verifies the first part of the theorem. In addition, for (2), components $h_i^{(1)}$ and $h_i^{(2)}$ are related by an increasing monotonous function in such a way that the index of the function that produces the minimum in $\{h_i^{(1)}\}$ is the same one that produces the minimum in $\{h_i^{(2)}\}$. This proves the second part.



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-Casermeyro ^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 xxx

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-Casermeyro).

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.

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

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:

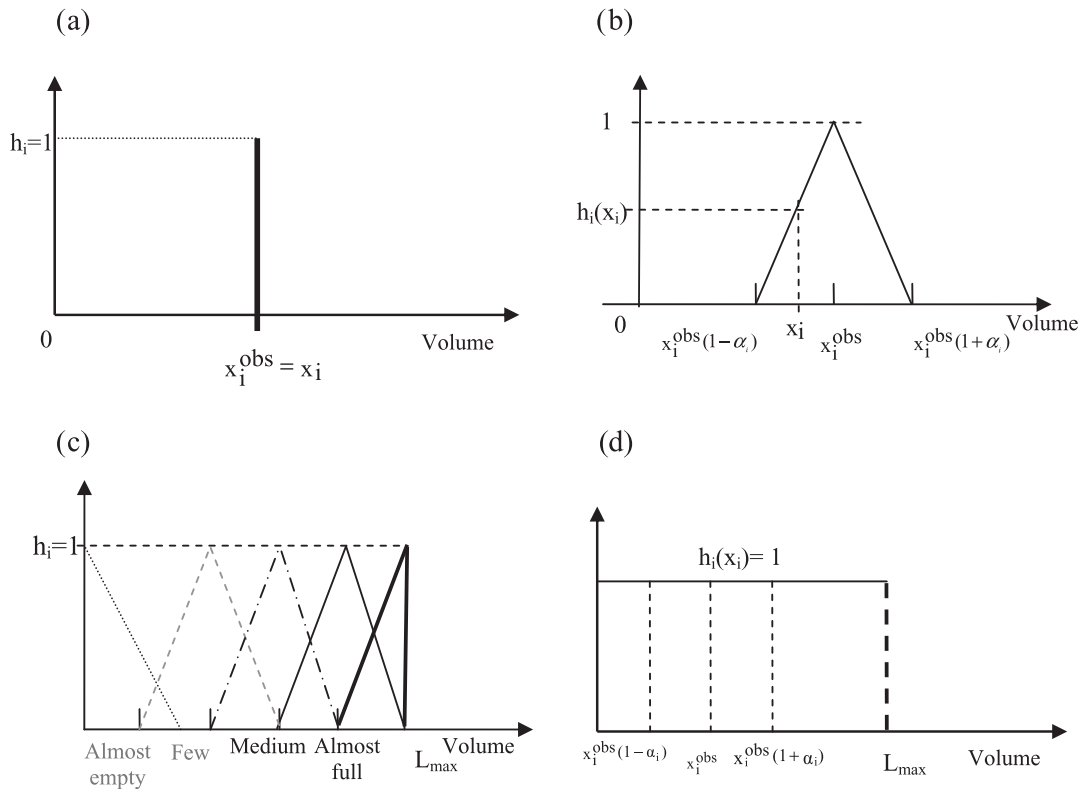


Fig. 1. Membership functions for (a) fix number, (b) crisp number, (c) fuzzy information and (d) missing value.

- $x_i^{obs} - \alpha_i x_i^{obs} \leq x_i \leq x_i^{obs} + \alpha_i x_i^{obs}$, 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 .
- 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 $h = \min(h_i)$.

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

subject to

- Constraints related to the membership functions:

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

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:

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

where N is the number of transit stops

- Constraints related to vehicle conditions:

$$l_k \geq 0 \text{ and } l_k \leq L_{\max} \tag{5}$$

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.

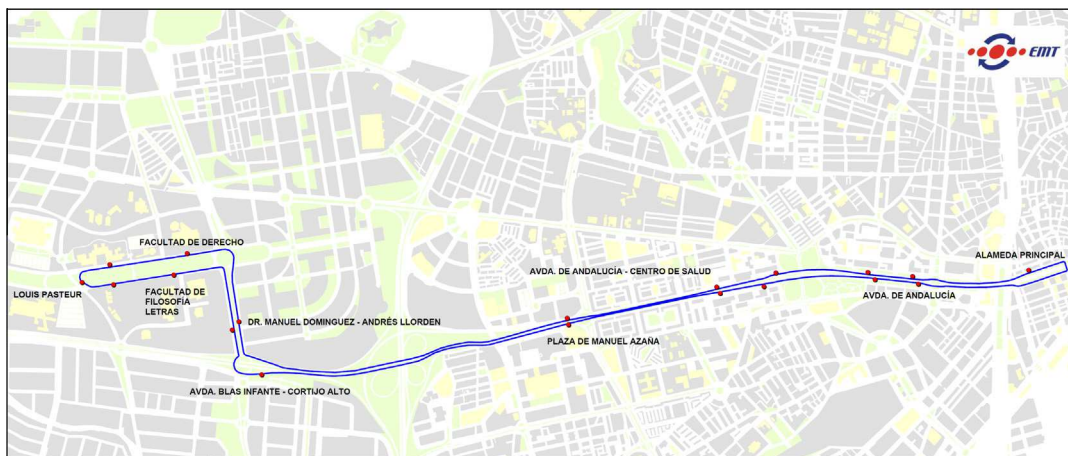


Fig. 2. Example of a transit line in Malaga.

Table 1
Alightings 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.

$$\epsilon = \frac{\sum_{i=0}^n |x_i - x_i^{true}|}{n} \tag{8}$$

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

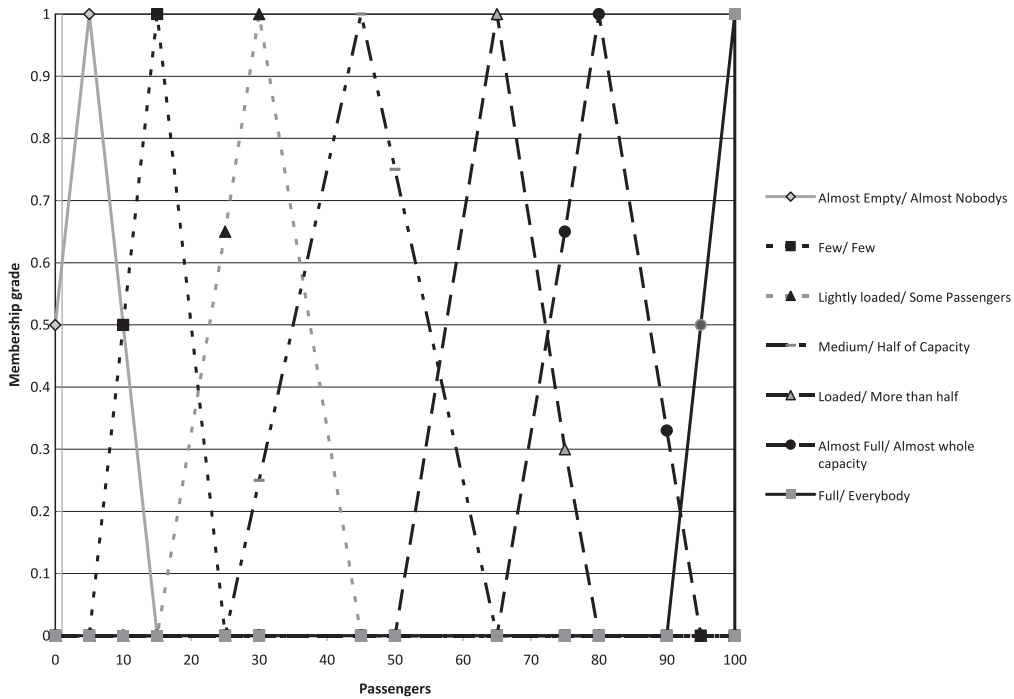


Fig. 3. Membership functions of loads and alightings in a transit line.

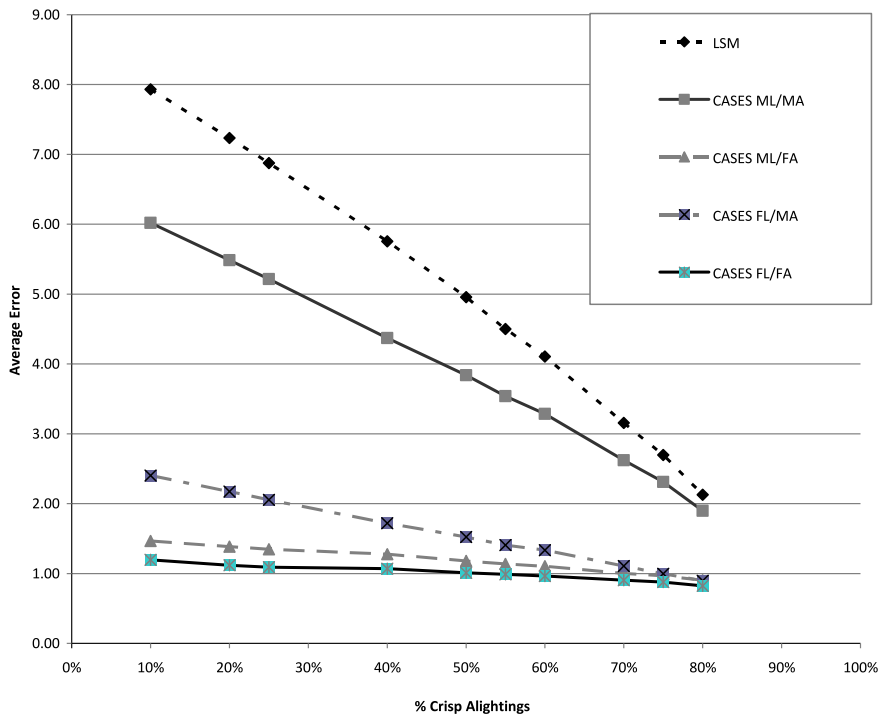


Fig. 4. Average error evolution in the case of a transit line in Malaga.

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 used 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-Casermeyro, 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.