Journal of Transport Geography 19 (2011) 1081–1092

Contents lists available at ScienceDirect

Journal of Transport Geography

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

Transit ridership forecasting at station level: an approach based on distance-decay weighted regression

Javier Gutiérrez [a,*], Osvaldo Daniel Cardozo [b], Juan Carlos García-Palomares [a]

[a] Departamento de Geografía Humana, Universidad Complutense de Madrid, C/Profesor Aranguren, s/n, 28040 Madrid, Spain
[b] Departamento de Geografía, Universidad Nacional del Nordeste, Av. Las Heras 727, Resistencia, Argentina

ARTICLE INFO

Keywords:
Distance-decay weighted regression
Direct ridership models at station level
Metro
GIS
Madrid

ABSTRACT

This article develops a rapid response ridership forecast model, based on the combined use of Geographic Information Systems (GIS), distance-decay functions and multiple regression models. The number of passengers boarding at each station in the Madrid Metro network is estimated as a function of the characteristics of the stations (type, number of lines, accessibility within the network, etc.) and of the areas they serve (population and employment characteristics, land-use mix, street density, presence of feeder modes, etc.). The paper considers the need to evaluate the distance threshold used (not the choice of a fixed distance threshold by assimilation from other studies), the distance calculation procedure (network distance versus straight-line distance) and, above all, the use of distance-decay weighted regression (so that the data from the bands nearer the stations have a greater weighting in the model than those farther away). Analyses carried out show that weighting the variables according to the distance-decay functions provides systematically better results. The choice of distance threshold also significantly improves outcomes. When an all-or-nothing function is used, the way the service area is calculated (straight-line or network distances) does not seem to have a decisive influence on the results. However, it seems to be more influential when distance-decay weighting is used.

1. Introduction

The usefulness of Geographic Information Systems (GIS) in transport planning has widely recognized (Nyerges, 1995; Miller, 1999; Miller and Shaw, 2001). GIS provide a flexible framework for delineating service areas and determining the population and employment covered within each area (O'Neill et al., 1992; Peng and Dueker, 1995; Hsiao et al., 1997; Murray et al., 1998; Murray, 2001; Zhao et al., 2003; Horner and Murray, 2004; Gutiérrez and García-Palomares, 2008). Coverage analysis offers planners useful information, since the number of people/employment served is a proxy for potential public transport demand, although it does not in itself provide an estimation of transit ridership. However, the joint use of GIS and multiple regression analysis does allow direct estimation models to be built which can forecast transit ridership at station level. Basically, this involves adjusting a multiple regression model where the dependent variable is station ridership and the independent variables are the characteristics of the stations and of their service catchment areas delimited with GIS tools. Although these models do not benefit from the 'regional' approach of the four-step travel models, they do give an in-depth local analysis and are simpler, offering the possibility of an immediate response.

To date, very few studies have used GIS and multiple regression analysis for transit forecasting (some exceptions are Parsons Brinckerhoff, 1996; Walters and Cervero, 2003; Kuby et al., 2004; Chu, 2004). Proposed models use standard distance thresholds for delimiting service areas and do not take into account that the greater the distance to stops and stations within the catchment area, the less public transport tends to be used. They simply use a function that takes the discontinuous form of all or nothing (inside or outside the service area). This article, in contrast, suggests treating station accessibility in more precise detail, choosing the best distance threshold in terms of demand and using distance-decay functions within the station catchment areas. To do this, the existing relationship between the distance to the station and the number of passengers entering on foot is analyzed. This relationship is used as a weight for the explanatory variables in a multiple regression model, in order to obtain a better estimation of the number of passengers entering stations in the Madrid Metro network.[1] This is what we call a distance-decay weighted regression model. Once the model is adjusted, it can be used to forecast boardings for future stations.

* Corresponding author.
E-mail addresses: javiergutierrez@ghis.ucm.es (J. Gutiérrez), osvaldodcardozo@yahoo.com.ar (O.D. Cardozo), jcgarcia@ghis.ucm.es (J.C. García-Palomares).

[1] The Madrid Metro network, at 215 km long, with a total of 191 stations and 618 million trips per year, is the one of the largest in Europe.

0966-6923/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtrangeo.2011.05.004

The article has five sections. After this brief introduction, Section 2 gives a review of the background to direct estimation models of demand at station level and transit accessibility. Section 3 proposes a set of methodological improvements related to calculating the distance to stations and adjusting distance-decay functions. Section 4 shows the regression model fit and discusses the results. Section 5 contains some final remarks on the methodology used, the model obtained and its possible application.

2. Background

Light rail transit is undergoing a resurgence, both in North America and Europe. Many cities in the United States have returned to light-rail transit with hopes of coaxing drivers out of their cars in congested corridors (Kuby et al., 2004). Critics of light rail argue that North-American cities have neither enough jobs downtown nor enough residents along their corridors to generate substantial ridership. However, ridership can be increased by transit-oriented developments (TOD) (see Curtis et al., 2009). By siting housing, workplaces and other urban activities within an easy walk to rail stations, proponents maintain that transit and walk trips will substitute what would otherwise be private-car travel (Cervero, 2004). Direct models are especially appropriate for forecasting ridership at station level along transit corridors and for estimating the travel impacts of transit-oriented developments.

2.1. Direct ridership models

Transit agencies need to know the impacts of changes to their network on transit ridership.[2] In order to forecast public transport demand, most planners use regional four-step travel forecasting models that consider trip generation, distribution, mode choice and route assignment (see, for example, McNally, 2000). In theory, the complexity of these models makes them the best tools for evaluating new transit facilities, but in practice there are several potential problems (Marshall and Grady, 2006), such as model accuracy (in most regional models much more attention has been paid to matching traffic counts on individual roadway segments than on matching transit loadings on individual route segments or stations), travel input data (estimation is typically based on relatively old household surveys, which may include only a small number of transit trips in the area of interest), sensitivity to land use (regional models are generally insensitive to land use), institutional barriers (transit providers are often not part of the modeling process) and cost of use (four-step travel models are cumbersome and expensive).

A need has emerged for simpler models that are capable of generating demand estimates quickly and economically. Direct models based on multiple regression analysis are a complementary approach to estimating ridership as a function of station environment and transit services features (Kuby et al., 2004; Chu, 2004; Cervero, 2006). Such models are a quick-response and less expensive alternative to the four-step travel forecasting models. They also capture better the influences on travel demand of the built environment of the station. Traditional regional demand models are sensitive to changes in land use at meso and macro scales, but their resolution tends to be too coarse to pick up fine-grained design and land-use mix features of neighbourhood-scale initiatives such as new urban planning and transit-oriented developments. Direct models lack the regional perspective of four-step models, but they are directly responsive to land-use characteristics within the station catchment areas (Cervero, 2006).

Chu (2004) argues that the traditional four-step process is ineffective for assessing the impacts of changes to the network or service levels on transit patronage. The reasons for this have to do with accuracy and relevancy. The four-step process was designed as a planning tool for large-scale capital intensive changes. The unit of analysis is typically at the zonal level. Errors from this process may be larger than the impacts of low-cost operational changes. Besides accuracy and relevancy, the traditional four-step process lacks flexibility in the possible changes to a transit network or services offered that a transit agency can explore and implement.

Direct ridership models based on the use of regression analysis combine features from all four steps of the traditional travel model. It is clearly not as comprehensive and systematic as the four-step process. Yet its transparency and high explanation power may make it useful for experimenting with different alignments of new LRT corridors, for which the values of all independent variables could be easily calculated (Kuby et al., 2004).

Walters and Cervero (2003) highlight the ability of direct ridership models to: (i) address combinations of transit alignments, station locations and vehicle technology types, (ii) conduct quick-response evaluation of variations in parking, feeder bus service, station spacing and transit speed and frequency, (iii) capture effects of local land use characteristics, such as increased densities and improved walkability, within transit station areas and transit-served communities, (iv) consider competing or parallel transit services a single corridor, with their respective shares of overall transit ridership, and (v) be developed, calibrated, validated and applied within an accelerated time schedule. They do not claim that direct modeling can replace the more regional scale four-step models; combining the direct demand model with the conventional four-step travel models accounts for both macro travel patterns and micro-area sensitivities for multi-station transit concepts (Walters and Cervero, 2003).

2.2. Factors affecting ridership

From the point of view of direct estimation models at station level, factors affecting ridership can be classified into three types: built environment dimensions, socioeconomic factors and characteristics of the stations. The built environment is thought to influence travel demand along three principal dimensions: density, diversity, and design (Cervero and Kockelman, 1997). Urban density is the critical driver of transit ridership. The evidence for a positive relationship between population density and transit ridership is well established at station level (Seskin and Cervero, 1996). The significance of urban density is that the more people living and/or working in close proximity to transit, the greater the likelihood the service will be used (Murray et al., 1998). Land-use type and mix also influence transit use, although less so than density (Parsons Brinckerhoff, 1996). Land-use mix (diversity) produces a more balanced demand for public transport over time (reducing differences between peak and off-peak periods) and in space (in terms of direction of flow) (Cervero, 2004). Filion (2001) found that mixed-use suburban centers have been successful in attaining higher transit use than the typical suburban area. Moreover, neighbourhoods that are more walkable (design) favor access to stations on foot and increase transit ridership (Cervero, 2002). Traditional grid street patterns (small street blocks and well-connected roads) enhance pedestrian access to public transport facilities. In recent years, many comparative studies of land use and transportation have responded to this problem by promoting neo-traditionalism, with its emphasis on transit-oriented urban design (Hsiao et al., 1997; Loutzenheiser, 1997).

[2] Although demand forecasting is a critical issue in the planning process of new stations, other criteria must also be taken into account, such as social exclusion. Station locations which have high population densities but low rates of economic activity should not be directly excluded as candidates if the models do not predict a high number of riders for them.

Population and employment within catchment areas (density) are important factors for estimating ridership, but more or fewer riders can be expected according to the type of population (or employment) in the catchment area, since the correlation of transit ridership with socioeconomic variables, such as household income, age, race/ethnicity and car ownership, is well known (see, for example, Cristaldi, 2005; Giuliano, 2003; Cervero and Duncan, 2002). Thus, for example, an increase in real per capita income and car ownership is associated with patronage decline (Gómez-Ibáñez, 1996; Wachs, 1989; Kitamura, 1989). Ethnic/racial minorities and immigrants are substantially more likely to use transit (Jin, 2005).

The characteristics of the stations are also relevant for explaining ridership. The number of riders is related to the type of station (intermediate, terminal, interchange or intermodal). Terminal stations are the nearest stations for residents of a large area beyond the end of the line and people are willing to walk longer to reach this type of station (O'Sullivan and Morrall, 1996). Interchange stations are more attractive for travellers than intermediate stations and tend to capture more riders, while intermodal stations also tend to have higher boardings, since they receive riders from other transport modes. Station spacing influences the size of the catchment areas and, indirectly, ridership; additional spacing around a station would draw additional riders to that station (Kuby et al., 2004). The distance from the station to the central business district – CBD (Pushkarev and Zupan, 1982) or its centrality within the network (Kuby et al., 2004) are also relevant, since people tend to use public transport more frequently in central areas than in peripheral ones. The presence of feeder modes (bus stops) near the station and park-and-ride facilities can also significantly affect ridership, since some riders reach the station by public transport or by car (Kuby et al., 2004). Service frequency influences patronage, but the use of this variable as a predictor can produce endogeneity problems between service supply and demand; while transit use is, to a large extent, a function of transit service supply, transit service supply is, of course, largely a function of transit demand (Taylor and Fink, 2003).

Direct ridership models use these types of variable as predictors at station level. They estimate ridership as a function of station environments and transit service features, using multiple regression because of its ability to simultaneously evaluate the effects of a large number of factors. Multiple regression is flexible, widely used and easily understood by a broad audience (Kuby et al., 2004). In addition, the resulting model can be used for predictive purposes. It relies on transit stations as the unit of analysis, but cumulating forecasts for individual stations then yield overall ridership estimates for a planned corridor (Walters and Cervero, 2003). The few direct ridership models proposed so far have been tested in US or Canadian cities. Table 1 summarizes the independent variables included in these pioneer works. Some of these models (Parsons Brinckerhoff, 1996; Kuby et al., 2004) fit data from several cities and include city-wide variables, such as heating and cooling degree days or primary metropolitan statistical area populations. Others consider only data of a particular study area (Walters and Cervero, 2003; Chu, 2004) and, logically, do not use city-wide variables.

The Parsons Brinckerhoff (1996) study used a cross-sectional regression analysis of station-level data in order to forecast daily station boardings on a new rail line in Charlotte-Mecklenburg, North Carolina. It combined data from different light-rail transit systems (261 stations on 19 lines in 11 US and Canadian regions). As predictors, the multiple regression model used density variables (population and employment) and the characteristics of the stations (for example, distance to the CBD, park-and-ride facilities and feeder-bus services). Only the presence of park-and-ride facilities was not significant at the 0.01 level.

Kuby et al. (2004) used a multiple regression model to determine factors that contribute to higher light-rail ridership, pooling stations from nine different light rail transit systems. Employment, population and renters within walking distance were significant, as were bus lines, park-and-ride facilities and centrality. Dummy variables for terminal and transfer stations and international borders, as well as citywide variables (such as heating and cooling degree days or metropolitan area population) were also positive and significant. Only two variables included in the final model were not significant at the 0.05 level (stations serving airport passenger terminals and the proportion of metropolitan area employment covered by the network). All the variables had the hypothesized positive or negative coefficients. The model had an R^2 value of 0.72, higher than the R^2 value of 0.53 obtained in the Parsons Brinckerhoff study that analyzed only non-CBD stations. The authors concluded that the resulting model may be useful as a first-cut one-step approach for predicting demand for possible light-rail alignments.

Walters and Cervero (2003) (see also Cervero, 2006) developed a direct ridership model in order to generate ridership forecasts for alternative light rail and heavy rail extensions to the San Francisco Bay Area Rapid Transit (BART) system and to guide future land-use policies for the affected corridor. Two final equations with logical and interpretable combinations of variables and the highest explanatory powers were selected. Both equations include variables such as population–employment densities around stations, transit technologies (heavy rail versus commuter rail technology), train frequencies and catchment population. The first equation adds variables related to access and egress (parking supplies and feeder-service levels). The two equations explain between 87% and 90% of the variation in station ridership. In addition, the direct-modeling approach made it possible to select and introduce variables that helped policy-relevant sensitivity tests to be conducted. It was thus proved, for example, that concentrated land development around stations can yield significant ridership benefits.

Chu (2004) developed a direct model using a Poisson regression in order to explain the ridership of transit stops in Jacksonville, Florida. Variables measuring the characteristics of the catchment areas, the pedestrian environment, accessibility to population and employment, interactions with other modes and competition from other stops in the catchment areas all played a statistically significant role in stop patronage for average weekdays. All these variables had the expected coefficient signs. The presence of these variables improved the log likelihood value by 54%.

2.3. Distance decay of ridership

A consideration of the impact of service area characteristics on the demand for public transport implies recognition of the importance of accessibility. According to Bertolini (1999), accessibility is not just a feature of a transportation node ('how many destinations, within which time and with which ease can be reached from an area?'), but also of place (in our case, 'with which ease can the station be reached?'). It is well-known that walking distance has a negative impact on transit use. Keijer and Rietveld (2000) found that in the Netherlands people living in a buffer 500–1000 m from a railway station tend to use rail services 20% less than people living less than 500 m from stations. Untermann (1984) found a distance-decay relationship in which most people were willing to walk 500 ft, 40% would walk 1000 ft, but only 10% would walk a half mile. Zhao et al. (2003) conducted an onboard transit survey to determine the effect of walking distance on transit use, showing that transit use deteriorates exponentially with walking distance to transit stops. Levinson and Brown-West (1984) classified bus riders by walking distance and car ownership rates, and compared them to the number of dwelling units in each status. A series of ridership penetration curves were calculated, showing that patronage declines linearly with increasing walking distance.

Accessibility has been recognized as one of the most important factors affecting transit use. In fact, rail travel levels are high if both

Table 1
Independent and dependent variables used in previous regression models for estimating transit ridership at the station-level.

Author	Dependent variable	Independent variables (included in the final model)	Transit mode
Parsons Brinckerhoff (1996)	Daily station boardings	Population density: natural log of persons per gross acre within 1/2 mile of station Employment density: (employees per gross acre within 1/2 mile of station) × (natural log of employees/1000) Terminal station (0 = no, 1 = yes) Park-and-ride (0 = no, 1 = yes) Feeder bus services (0 = no, 1 = yes) Catchment size: natural log of distance to nearest adjacent station Distance to CBD: natural log of miles between station and CBD along shortest light-rail route	Light-rail transit (11 US and 2 Canadian regions)
Kuby et al. (2004)	Average weekday boardings	Employment within walking distance Population within walking distance Stations that serve airport passenger terminals International border (station location) Park-and-ride spaces Bus connections: number of different bus lines intersecting with a station Heating and cooling degree-days Terminal station Designated transfer station Normalized accessibility Percentage of PMSA employment covered by system Percent renters within walking distance	Light-rail transit (nine US cities)
Walters and Cervero (2003)	AM peak period entrances and exits	Population–employment densities around stations Transit technologies (light-rail versus heavy-rail) Train frequencies Catchment population Parking supplies Feeder service levels	Light-rail and heavy-rail (San Francisco, California)
Chu (2004)	Weekday boarding	Median household income (000s) in catchment area Jobs in catchment area by road 0-vehicle households in catchment area Share of persons under 18 (0–1) in catchment area Share of persons 18–64 (0–1) in catchment area Share of persons female (0–1) in catchment area Share of persons Hispanic (0–1) in catchment area Share of persons White (0–1) in catchment area Transit Level of Service (TLOS) within 1-min walking (0–100) Transit stops within 2–5 min walking (0–100) Pedestrian factor (0–1) Persons up and downstream without transfer (000s) in 1 h Jobs up and downstream without transfer (000s) in 1 h Including a trolley stop (1 if present; 0 otherwise) Number of other TLOS stops in catchment area	Transit bus (Jacksonville, Florida, USA)

the origins and destinations are in close proximity to a station (Cervero, 1994). Therefore, concentrating housing and employment within several hundred feet of a rail station will produce far more riders than placing the same level of development a half mile away (Bernick and Cervero, 1997). Regional models are non-sensitive to these variations in land use, as they assume that all trips in any zone start from its centroid. In contrast, direct ridership models based on GIS tools can capture the tendency of patronage to decay with walking distance to a station. However, the few existing direct estimation models do not treat walking distance adequately, as they consider population or employment levels within the station catchment area from an all-or-nothing approach.[3] Thus, the Parsons Brinckerhoff study (1996) includes as predictors population and employment density within a 0.5 mile buffer. Walters and Cervero (2003) calculate total population and employment within several buffer distances (0–0.25, 0.25–0.5, 0–0.5 and 0–1 miles), but the best-fitting multiple regression equations include only total population and employment densities within a half-mile of stations. A half-mile walking distance was also selected by Kuby et al. (2004), but they calculated the shortest network paths (not Euclidean distances). A special raster-based approach was developed in this study to improve upon standard GIS buffering commands (Upchurch et al., 2004).

In sum, previous direct ridership models at station level used fixed distance thresholds (Euclidean or network). This means that they cannot reflect the greater tendency to use public transport when the distance within the station catchment area is shorter. It implies that they are not able to reflect the impact on travel of concentrating housing and employment at a longer/shorter distance from the station in cases where these developments are located within the station catchment area. To overcome this problem we propose the use of a distance-decay weighted regression, that is, the combined use of multiple regression models and distance-decay functions.

3. Data and methodology

3.1. Data

3.1.1. Data sources

The following data (all referring to 2004) have been used in order to build the ridership forecast model:

[3] Kuby et al. (2004, p. 244) wonder whether a more sophisticated distance assumption than all-or-nothing buffers would work even better. On the other hand, Cervero (2006, p. 289) claims that this type of model captures dynamics that bigger models miss, such as the tendency of patronage to decay exponentially with walking distance from a station. However, in praxis all the previous papers revised use functions of all-or-nothing.

- Metro boardings.
- The Transport Authority of Madrid (Consorcio Regional de Transportes de Madrid) provided boarding data for November 2004 at station level.
- Stations and stops in the public transport network.
- GIS layers containing Metro stations and bus stops were used in order to delineate Metro station service areas and consider feeder buses.
- Street network.
- This layer was essential for computing station service areas using network distances. It was also necessary for calculating urban design indices (street density).
- Socioeconomic variables.
- At transport zone (TZ) level, several socioeconomic variables provided by the 2004 mobility survey were available, such as population, employment, workers, students, foreigners, number of households, car ownership, etc.
- Origin of trips.
- The 2004 mobility survey provides a table showing *x*, *y* coordinates for the origin of trips accessing the Metro on foot. With these data it was possible to compute distance-decay functions (see Section 3.2).

3.1.2. Selection of variables

The dependent variable in multiple regression models is *monthly station ridership* (*Metro boardings*). The independent variables initially selected for their influence on transit ridership (also referring to 2004) are the following.

3.1.2.1. Service area characteristics. Station ridership is influenced by population and employment within the catchment area and their characteristics. For population variables, instead of working with the *total population* of each station service area, population was segmented into several groups: *workers*, *foreigners*, *population under 20 years old*, *population over 60 years old* and *non-car owning households* (household income data were not available). Likewise, *total employment* was divided into *commercial, administration, education, health* and *industrial sectors*, since some jobs (particularly those related to education and commerce) attract more trips than others.

We propose to measure *land-use mix* by using the reciprocal of the variation coefficient of the area covered by different land uses within the station service area (higher values indicate higher diversity of use). This measure is easily computable and interpretable. An urban design indicator has been calculated using the street network layer: *street density* within the station area (ratio between street length and service catchment area). Street density can be considered as an indicator of walkability (Zhu and Lee, 2008). It is hypothesized that land-use mix and street density favor transit use.

In order to take into account the importance of feeder buses on station ridership, two variables were considered: the number of *urban* and *suburban bus lines* with stops within the 200 m station catchment area, since these are also means of mass public transport and show evidence of interdependence with the Metro. Finally, *park-and-ride spaces* were treated as a dummy variable.

3.1.2.2. Station characteristics. *Terminal* and *intermodal* stations tend to have more riders than other stations, the former because they have bigger catchment areas and the latter because they receive riders from other transport modes. They have been treated as two separate 0–1 dummy variables. Interline *transfer* stations were also expected to have higher boardings. Instead of giving these stations a dummy value of 1 (Kuby et al., 2004), the *number of lines* passing through each station was considered (the more lines passing through a station, the more attractive it is).

To measure *nodal accessibility* (accessibility of the station within the network), the economic potential indicator (Hansen, 1959) was preferred to other centrality measures, such as distance to the CBD (Parsons Brinckerhoff, 1996) or average travel times within the network (see Kuby et al., 2004), because of its gravitational nature, which gives more realistic results. The nodal accessibility of each station within the network was calculated according to the equation:

$$A_i = \sum_{j=1}^{n} \frac{E_j}{C_{ij}}$$

where A_i is the accessibility of station *i*, E_j is the employment within the service area of destination station *j*, and C_{ij} is the cost (travel time) through the Metro network between stations *i* and *j*.

In order to prevent endogeneity problems in the model, *service levels* were not considered. Service levels influence station ridership, but the dominant causality is that lines crossing central and highly populated areas are provided with better service levels.

3.2. Methodology: service areas and distance-decay functions

Cervero (2006) notes that direct models should capture the tendency for patronage to decay exponentially with walking distance to a station. Nevertheless, the direct estimation models proposed so far only delimit service areas and calculate their characteristics to obtain predictors. In order to capture the tendency of patronage decay with walking distance to a station, three basic factors have to be taken into account: the procedure for calculating walking distance, the choice of service area distance threshold and consideration of a distance-decay function.

3.2.1. Calculation procedure for the walking distance to transit facilities

Direct estimation models based on GIS tools and regression analysis have followed the coverage analysis tradition: they delimit transit facility service areas from a certain straight-line distance (GIS buffering commands).[4] Only Kuby et al. (2004) calculate network distance based on a raster approach (for more details see Upchurch et al. 2004).

Euclidean-distance calculation is an unrealistic approach when considering walking distance to transit facilities, since a pedestrian follows the layout of the street network, not a straight line, in order to access stations. In practice, the distance calculation procedure is important since a relationship exists between distance measure methods and ridership. Gutiérrez and García-Palomares (2008) calculated coefficients of determination between metro ridership at the station level and the population covered according to two distance methods, concluding that the network distance method provides better estimates of metro ridership than the Euclidean distance method. Following Gutiérrez and García-Palomares, our direct estimation model is based on calculating distances across the network to delimit station service areas. Unlike Kuby et al. (2004) we use a vector GIS for network distance calculations, which provides more precise data than that obtained with a raster approach.

[4] Most of the coverage analysis delimits service areas using Euclidean distances. Some exceptions are O'Neil et al. (1992), Hsiao et al. (1997), Horner and Murray (2004) and Gutiérrez and García-Palomares (2008), who calculate distances along the street network, simulating real routes followed by the population on their way to transit facilities. More accurate results are obtained using distances along the network than Euclidean distances, since the buffer method tends to overestimate the area and the population covered by the public transport network.

3.2.2. Choice of service area distance threshold

The standard walking distances used to delimit service areas in most transit research are 0.25 miles (400 m) for bus stops and 0.5 miles (800 m) for rail stations (see, for example, ÓNeill et al.; 1992; Hsiao et al.; 1997; Murray, 2001; Zhao et al., 2003; Kuby et al., 2004). This represents the maximum distance that most people are willing to walk to use transit. However, this distance threshold should not be considered as a hard standard, since it is area specific (urban bus networks in different cities may have different critical points in the deterioration of transit use due to increasing walking distance) and mode specific (people are willing to walk farther to access a network with more spaced out stations than to a network with a greater density of stations) (El-Geneidy et al., 2010).

The choice of critical distance for delimiting service areas is important in any ridership study and must be justified in terms of demand. With a very high distance threshold, areas far from the station are taken into account, which add few users and can distort the final results. On the other hand, a very low distance threshold creates very small service areas, which leave out most of the station's riders. In theory, station demand data could be analyzed to find the critical point from which demand is irrelevant. However, this critical point might not really exist; there may be just a gradual fall in demand. This study proposes choosing the distance threshold based on the correlation existing between independent variables and ridership. The threshold chosen will be the one for which maximum correlation is obtained, i.e., the threshold with the greatest explanatory capacity of the independent variables.

3.2.3. Distance-decay functions

Walking distance to transit facilities is an important factor of transit ridership: the farther away people are from transit facilities, the less likely it is they will use transit (see Cervero, 2004, or Cervero and Duncan, 2002). Since the tendency to use public transport decreases as the distance to transit facilities increases, ridership models should be able to reflect this trend. However, the few existing direct estimation models merely consider one or two distance thresholds (in general, a single band, taking the standard 0.25 or 0.5 miles as threshold), ignoring that within each band the tendency to use transit is greater near the station than at the outer edge. This article presents a distance-decay weighting (Fig. 1), such that when calculating variables around stations, a greater weight is assigned to areas that are nearer the stations (where there is a higher probability of using public transport) than to those that are farther away (where the probability is lower). Only models that include a distance-decay function are sensitive to the different distributions of independent variables within the station service area.

3.3. Calibration of distance-decay functions

In order to calibrate distance-decay functions, spatially disaggregated data on public transport use are needed. In the case of the Madrid Metro, the 2004 mobility survey provides geographical coordinates for the origin of recorded trips. Around 17,000 trips accessing stations on foot were selected, distinguishing between those starting from home and those starting elsewhere. It is worth noting that 80% of passengers entering the Madrid Metro network do so on foot and only 20% come by other modes of transport.

To estimate trip ratios according to station distance, several 100 m-wide bands were created around stations through the street network, up to a maximum limit of 1500 m. In order to take competition between stations into account, Thiessen polygons were generated. For each of the service areas, the population and employment covered by each distance band were calculated. The areal interpolation method (O'Neill et al., 1992; Chakraborty and Armstrong, 1997) was used[5] according to the equation:

$$P = \sum_{i=1}^{n} P_i * ap_i \qquad (1)$$

where P is the population in the service area of the metro station, $i \ldots n$ the transport zones totally or partially covered by the service area, p_i the Population in transport zone i, and ap_i is the area proportion of the transport zone i that is contained within the service area

This method was applied not for the service area as a whole but band by band.

For those travellers accessing the Metro network from home, ratios have been calculated by dividing these trips by the population residing within the corresponding distance band. For those accessing it from elsewhere, ratios were calculated by dividing the trips by the employment located in the corresponding distance band. It is clearly apparent that both the trips/population and the trips/employment ratios have a decreasing trend with distance (Figs. 2 and 3). Thus, for instance, in the 0–100 m band approximately 0.50 trips per person and day are registered, but this value falls to 0.30 in the 500–600 m band. The fit function is different for population (linear function) than for economic activity (exponential function), but in both cases the determination coefficients obtained (r^2) are very high (0.972 and 0.952), proving that a close correlation exists between distance to stations and use of the Metro.

Both these distance decay functions were used to weight attributes at station surroundings for each distance band, the first for population-related variables and the second for employment-related variables.

4. Multiple regression models

4.1. Distance-decay weighted regression

To explore the influence of every single predictor on ridership and to test for multi-collinearity among independent variables, bivariate correlations among the selected variables were calculated.[6] It was hypothesized that the ridership of intermodal stations could not be explained adequately by a direct model, as access is mainly via transport modes other than pedestrian (so demand is much higher than expected and varies significantly from station to station). Because of this, two matrices of correlation coefficients among variables were calculated, the first matrix including all stations (Table 2) and the second excluding intermodal stations (Table 3).[7] In both tables the service area variables were obtained using network distances and distance-decay weighting, taking a threshold distance of 800 m as the service limit area (see Section 4.2).

Correlation coefficients between ridership and independent variables (and their significance) tend to be higher in Table 3, particularly in the case of the service area variables. When intermodal stations were eliminated, the correlation coefficients of all the variables selected were found to be significant at a significance level of 0.05, with the exception of parking spaces, terminal stations,

[5] Since transport zones were delineated taking into account homogenous land use and building, employment and residential distributions are rather uniform within each transport zone. Under these circumstances, areal interpolation provides a good estimate for employment and population distributions within each service area.

[6] The most popular station (Sol) in Madrid, which is similar to Times Square in the New York Subway network, has not been considered because its high demand cannot be explained in quantitative terms. It is the most important attraction point in the mental map of people in Madrid who want to go to the city center for shopping or leisure purposes, particularly in the cases of tourists and immigrants.

[7] Following analysis of the normality of the independent variables, some of these were logarithmically transformed (by their natural logarithm), but in the end we used the original variables because we obtained better determination coefficients in the multiple regression models with non-transformed variables.

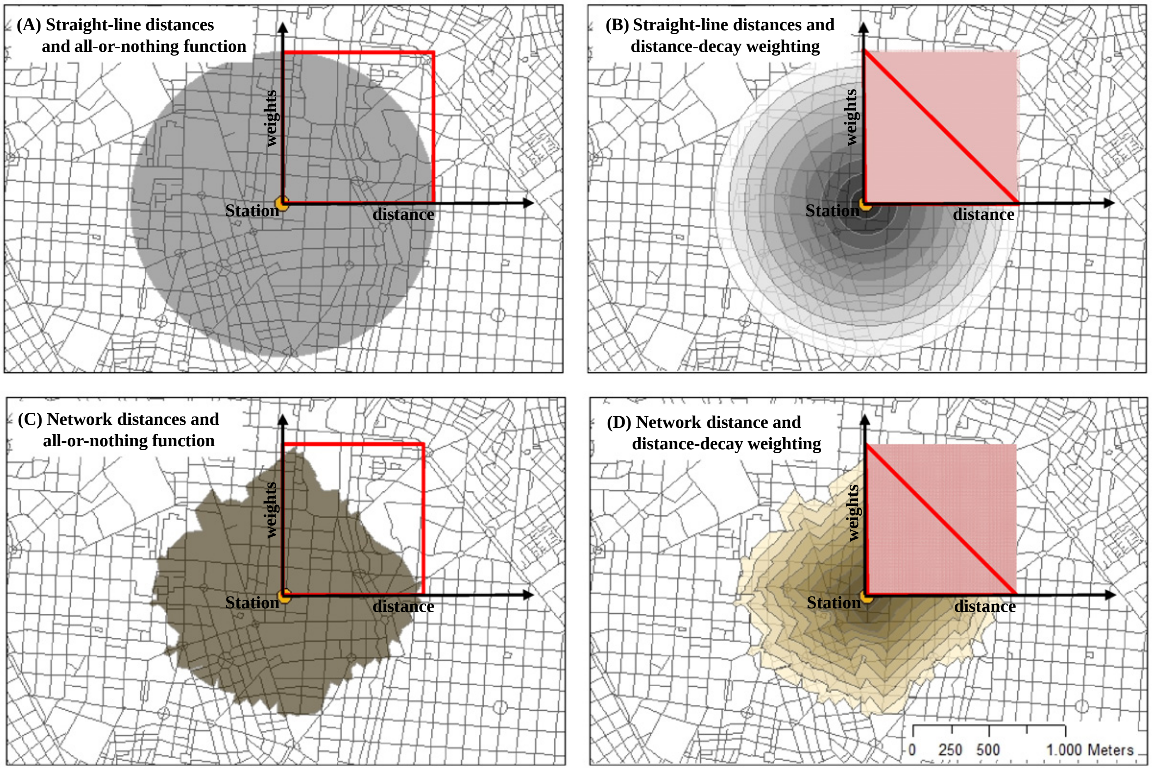
(A) Straight-line distances and all-or-nothing function
weights
Station
distance
(B) Straight-line distances and distance-decay weighting
weights
Station
distance
(C) Network distances and all-or-nothing function
weights
Station
distance
(D) Network distance and distance-decay weighting
weights
Station
distance
0 250 500 1.000 Meters

Fig. 1. Methods for calculating variables around stations by GIS, combining distance calculation procedures (straight line versus network distances) and distance-decay functions (all-or-nothing versus distance decay weighting).

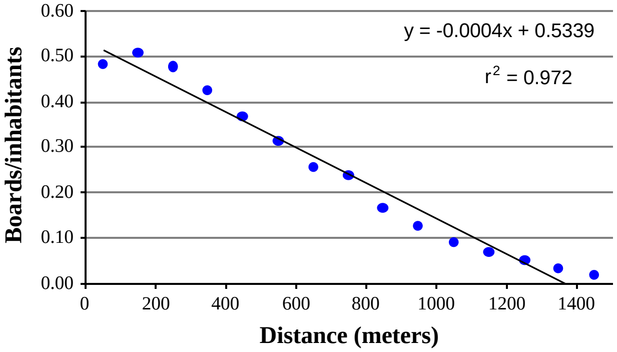

Fig. 2. Ratio of daily boards on the Madrid Metro and population (per distance bands).

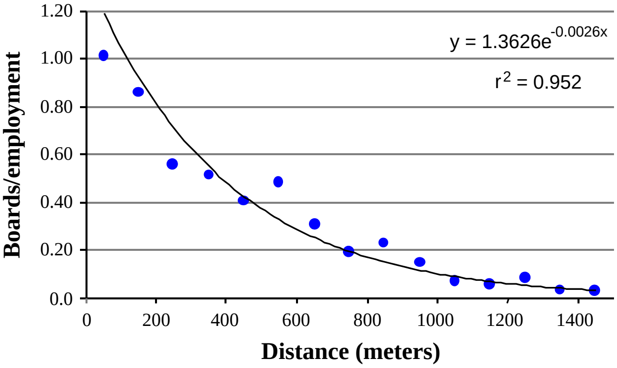

Fig. 3. Ratio of daily boards on the Madrid Metro and employment (per distance bands).

non-car households and the number of suburban bus lines within the catchment area. In general, employment variables in the catchment area are more explanatory than population variables. The numbers employed in the commercial sector (0.57) and in education (0.55) and the number of workers (0.55) are above the 0.50 threshold. There are also two variables related to station characteristics having high correlation coefficients: nodal accessibility (0.74) and the number of lines (0.66).

Correlations among employment variables are higher than those among population ones. The former range from 0.46 to 0.66 and the latter from 0.16 to 0.50 (Table 3). High correlation coefficients were also observed between the number of lines and nodal accessibility (0.66) (most of the transfer stations are located in central areas), people over 65 and foreigners (0.56) (both population groups tend to concentrate in central areas), and nodal accessibility and employment in administration (0.52) (these types of jobs are mainly located in central areas). All these variables were below the danger level of 0.7 (Clark and Hosking, 1986). The same trends are observed in Table 2.

A number of statistical analyses were run to test the different hypotheses and develop the best model for explaining and predicting boardings. They showed unexpected signs in some predictors, but signs, significance and determination coefficients tended to improve systematically when intermodal stations were eliminated. It is not surprising since most of the riders at this type of station come from transport modes other than pedestrian. In fact, their demand varies widely depending mostly on the amount of passengers provided by the feeder modes. They can be treated in a more realistic way using the four-step model (a regional approach) than by the inclusion of a simplistic dummy variable in a multiple regression model.

Thus, it was decided to eliminate intermodal stations in order to build a simple model that was intuitive and interpretable and had

Table 2
Matrix of bivariate correlations (r) between independent variables and boardings (all stations).

	Boardings	Intermodal (1 = yes; 0 = not)	Parking (1 = yes; 0 = not)	Nodal accessibility	Number of lines	Terminal (1 = yes; 0 = not)	Foreign population	Population under 20	Population over 65	Workers	Non car households	Employment in industrial sector	Employment in commercial sector	Employment in health sector	Employment in education sector	Employment in administration sector	Land use mix	Street density	Urban bus lines	Suburbanbus lines
Boardings	1																			
Intermodal	0.282[**]	1																		
Parking	0.012	0.316[**]	1																	
Nodal accessibility	0.685[**]	−0.020	−0.150[*]	1																
Number of lines	0.692[**]	0.077	−0.105	0.685[**]	1															
Terminal	−0.099	−0.069	0.041	−0.117	−0.089	1														
Foreign population	0.049	−0.050	−0.210[**]	0.029	0.010	−0.095	1													
Population under 20	−0.075	−0.080	−0.240[**]	−0.221[**]	−0.122	0.041	0.414[**]	1												
Population over 65	0.156[*]	−0.156[*]	−0.217[**]	0.066	0.058	−0.067	0.559[**]	0.493[**]	1											
Workers	0.393[**]	−0.003	−0.220[**]	0.333[**]	0.228[**]	−0.080	0.261[**]	0.273[**]	0.471[**]	1										
Non-car-households	0.055	0.023	−0.091	0.027	0.073	−0.103	0.261[**]	0.268[**]	0.404[**]	0.173[*]	1									
Employment in industrial sector	0.381[**]	−0.069	−0.210[**]	0.477[**]	0.300[**]	−0.142	0.146[*]	−0.032	0.188[**]	0.534[**]	0.032	1								
Employment in commercial sector	0.404[**]	−0.015	−0.220[**]	0.487[**]	0.399[**]	−0.129	0.170[*]	0.053	0.191[**]	0.440[**]	0.103	0.609[**]	1							
Employment in health sector	0.340[**]	−0.093	−0.190[**]	0.419[**]	0.283[**]	−0.096	0.088	0.015	0.157[*]	0.542[**]	−0.011	0.524[**]	0.468[**]	1						
Employment in education sector	0.329[**]	−0.121	−0.240[**]	0.457[**]	0.314[**]	−0.089	0.121	0.136	0.240[**]	0.562[**]	0.152[*]	0.597[**]	0.527[**]	0.506[**]	1					
Employment in administration sector	0.405[**]	−0.034	−0.183[*]	0.569[**]	0.380[**]	−0.115	0.009	−0.128	−0.010	0.423[**]	0.000	0.690[**]	0.562[**]	0.577[**]	0.625[**]	1				
Land use mix	−0.061	−0.056	−0.133	−0.238[**]	−0.135	0.103	0.273[**]	0.579[**]	0.407[**]	0.157[*]	0.160[*]	−0.044	−0.023	−0.064	0.035	−0.207[**]	1			
Street density	0.225[**]	−0.213[**]	−0.240[**]	0.287[**]	0.199[**]	−0.145[*]	0.445[**]	0.118	0.532[**]	0.365[**]	0.268[**]	0.373[**]	0.277[**]	0.254[**]	0.288[**]	0.267[**]	0.108	1		
Urban bus lines	0.452[**]	−0.004	−0.098	0.340[**]	0.271[**]	0.034	0.038	−0.040	0.261[**]	0.458[**]	0.107	0.408[**]	0.346[**]	0.379[**]	0.343[**]	0.426[**]	−0.024	0.340[**]	1	
Suburban bus lines	0.270[**]	0.160[*]	0.088	−0.084	0.082	−0.038	0.039	0.244[**]	−0.020	0.028	0.051	−0.086	−0.110	0.131	−0.113	−0.152[*]	0.104[*]	−0.230[**]	−0.067	1

Number of stations: 190.

[*] Correlation is significant at the 0.05 level (2-tailed).

[**] Correlation is significant at the 0.01 level (2-tailed).

Table 3
Matrix bivariate correlations (r) between independent variables and boardings (without intermodal stations).

	Boardings	Parking (1 = yes; 0 = not)	Nodal accessibility	Number of lines	Terminal (1 = yes; 0 = not)	Foreign population	Population under 20	Population over 65	Workers	Non car households	Employment in industrial	Employment in commercial	Employment in health sector	Employment in education sector	Employment in administration sector	Land use mix	Street density	Urban bus lines	Suburbanbuslines
Boardings	1																		
Parking	−0.117	1																	
Nodal accessibility	0.741[**]	−0.131	1																
Number of lines	0.663[**]	−0.101	0.661[**]	1															
Terminal	−0.122	0.091	−0.136	−0.094	1														
Foreign population	0.277[**]	−0.160[*]	0.087	0.049	−0.109	1													
Population under 20	0.074	−0.175[*]	−0.229[**]	−0.094	0.038	0.366[**]	1												
Population over 65	0.358[**]	−0.169[*]	0.055	0.042	−0.079	0.566[**]	0.479[**]	1											
Workers	0.555[**]	−0.251[**]	0.343[**]	0.229[**]	−0.088	0.261[**]	0.285[**]	0.502[**]	1										
Non car households	0.132	−0.05.	0.060	0.079	−0.116	0.233[**]	0.267[**]	0.394[**]	0.162[*]	1									
Employment in industrial	0.474[**]	−0.180[*]	0.438[**]	0.219[**]	−0.160[*]	0.142	−0.066	0.179[*]	0.485[**]	0.039	1								
Employment in commercial	0.566[**]	−0.168[*]	0.466[**]	0.376[**]	−0.148	0.152	0.038	0.204[*]	0.443[**]	0.122	0.596[**]	1							
Employment in health	0.440[**]	−0.153	0.389[**]	0.237[**]	−0.104	0.057	−0.001	0.133	0.561[**]	−0.005	0.512[**]	0.463[**]	1						
Employment in education	0.547[**]	−0.184[*]	0.444[**]	0.282[**]	−0.108	0.093	0.088	0.199[*]	0.557[**]	0.135	0.577[**]	0.511[**]	0.502[**]	1					
Employment in administration	0.426[**]	−0.163[*]	0.523[**]	0.297[**]	−0.132	−0.004	−0.158[*]	−0.041	0.379[**]	0.001	0.666[**]	0.499[**]	0.570[**]	0.607[**]	1				
Land use mix	0.193[*]	−0.069	−0.258[**]	−0.132	0.108	0.251[**]	0.576[**]	0.425[**]	0.186[*]	0.196[*]	−0.060	−0.030	−0.050	0.022	−0.224[**]	1			
Street density	0.363[**]	−0.175[*]	0.260[**]	0.158[*]	−0.168[*]	0.460[**]	0.102	0.505[**]	0.375[**]	0.256[**]	0.341[**]	0.240[**]	0.226[**]	0.231[**]	0.235[**]	0.107	1		
Urban bus lines	0.450[**]	−0.184[*]	0.269[**]	0.150	0.050	0.099	0.017	0.272[**]	0.449[**]	0.072	0.396[**]	0.395[**]	0.414[**]	0.374[**]	0.437[**]	0.043	0.338[**]	1	
Suburban bus lines	−0.127	0.126	−0298[**]	−0.143	−0.008	−0.003	0.273[**]	−0.073	−0.037	0.096	−0.183[*]	−0.144	0.074	−0.146	−0.243[**]	0.180[*]	−0.310[**]	−0.240[**]	1

Number of stations: 158.

[*] Correlation is significant at the 0.05 level (2-tailed).

[**] Correlation is significant at the 0.01 level (2-tailed).

Table 4
Multiple regression model for non-intermodal stations. Dependent variable: monthly boardings.

Independent variables	*B*	Std. error	*t*	Beta	Sign.	Tolerance
Nodal accessibility	1.230	.273	4.500	.461	.000	0.261
Number of lines	51977.301	26979.693	1.957	.171	.050	0.216
Foreign population	41.676	14.717	2.832	.125	.005	0.863
Workers	16.491	6.859	2.404	.135	.017	0.539
Employment in commercial sector	52.388	23.415	2.237	.117	.027	0.619
Employment in educational sector	82.047	42.133	1.947	.108	.053	0.548
Land use mix	3.591	1.177	3.051	.144	.003	0.755
Urban bus lines	2379.995	698.841	3.406	.166	.001	0.708
Suburban bus lines	318.374	201.320	1.581	.073	.116	0.804
Parking (1 = yes; 0 = not)	68639.669	30305.569	2.265	.097	.025	0.921

Number of stations: 158.
Number of independent variables: 10.
Degrees of freedom: 146.
F: 44.53 (significance = 0.000).
R: 0.868.
R^2: 0.753.
R^2 adjusted: 0.736.

high predictive power. The variables that were clearly not significant in explaining average boardings were eliminated: terminals, non-car households, people under 20 years old, people over 60, employment in the industrial, health and administration sectors, and street density. The final model contains predictors that are relevant to ridership from both the point of view of theory and policy: station characteristics (nodal accessibility and number of lines), socioeconomic variables (foreign population, workers, number of jobs in commercial and educational sectors), land-use mix and bus feeders (urban and suburban lines) (Table 4). This model has an R^2 of 0.753 (adjusted $R^2 = 0.736$), and an *F*-statistic value of 44.53, significant at the 0.000 level. The model thus explains around 75% of the variance in boardings for all non-intermodal stations.[8]

All of these variables have the expected coefficient signs and all are significant at the 0.05 level, except for one predictor (suburban bus lines in the catchment area). The coefficients offer relevant information on the elasticities between passenger boardings and each of the predictor variables. This means, for example, that an increase of 52.4 journeys per month (on average) is expected for each new job in the commercial sector within the service area (see Table 4). In addition, the model is very sensitive to the location of new developments within the station catchment area. Thus, for instance, 100 new commercial jobs located in the 0–100 m band from a station would produce an increase of 10,950 journeys per month, but if the new jobs were located in the 700–800 m band this would be only 1780.[9]

To analyze whether the independent variables are highly intercorrelated (multicollinearity), the 'tolerance' associated with each predictor was computed (Table 4). The tolerance of x_i is defined as 1 minus the squared multiple correlation between that x_i and the remaining x variables. All tolerance values are above 0.2, so there is no high multicollinearity (particularly in service area variables, all above the 0.5 threshold). In sum, the final regression model retains a small set of modestly inter-correlated variables.

Employment accessibility is highly significant in our final model as a measure of reachable opportunities and centrality of the station within the network: Metro use tends to be higher in central areas (many opportunities reachable within a short travel time from each station) than in peripheral ones. Variables of employment in commercial and educational sectors, as expected, are highly significant. Population groups that use the Metro very frequently (workers and foreigners) were included in the final model. The number of non-car households is not highly significant, probably because the Madrid Metro network covers very dense areas where having a car is not really an advantage. The number of urban lines near the station is highly related to station boardings. Land-use mix is significant too, but street density does not significantly influence ridership, probably because this variable is encapsulated in the calculation of service areas using network distances.

4.2. Advantages of distance-decay weighted regression

In order to verify the advantages of processing station accessibility in more detail to estimate station boardings, a number of multiple regression analyses were performed using the proposed model (Table 4) but entering variations in the calculation of service area variables: the procedure for delineating service areas (network distances versus straight-line distances), weights of the bands (distance-decay versus all-or-nothing function) and distance thresholds of the service area (from 200 m to 1200 m). Table 5 shows the determination coefficients (R^2) obtained combining different procedures: Euclidean walking distance with (column B) and without (column A) distance-decay weighting, and walking distance through the street network with (column D) and without (column C) distance-decay weighting (see also Figs. 1 and 4).

The most relevant finding is that weighting according to the distance-decay functions (columns B and D in Table 5) systematically provides better results (see also Fig. 4). Distance-decay weighting increases determination coefficients (2.5% on average and 3.6% on the 800 m threshold) if catchment area variables are calculated by straight-line distance (differences between columns B and A), but even more so (4.2% and 4.5%) if network distances are used (differences between columns D and C) (see Table 6). As expected, the best improvement in determination coefficients (5.1% on average and 6.9% on the 800 m threshold) is to be found between the simplest methodological combination (column A) and the most sophisticated one (D) (see Table 6).

Determination coefficients tend to grow between the 200 m threshold and the 800 m one and remain more or less stable to the 1200 m threshold (in the case of distance-decay weighting, see lines B and D in Fig. 4), or even decrease (in the case of the all-or-nothing function, see lines A and C). Considering too many

[8] Different multiple linear regressions were adjusted considering the network distance procedure and the distance-decay functions obtained in Section 3.2, but varying the service area limit from 200 to 1500 m. The analysis was carried out with the SPSS v.17 statistics package. The best-fitting multiple regression model was that obtained for an 800 m distance threshold.

[9] Elasticity should be multiplied by the number of new jobs and weighted according to the distance-decay function shown in Fig. 3, the average of the first eight bands being equal to 1. In this case, the weights were 2.09 for the 0–100 m band and 0.34 for the 700–800 m band.

Table 5
Sensitivity of the model (determination coefficients R^2) to changes in the procedures for delimiting service areas (straight line versus network distances), distance-decay functions (all-or-nothing versus distance decay weighting) and distance thresholds (from 200 to 1200 m).

Distance band	A Straight line distance and all-or nothing function	B Straight line distance and distance-decay method	C Network distance and all-or-nothing function	D Network distance and distance-decay method
200	0.677	0.676	0.669	0.663
300	0.680	0.689	0.663	0.695
400	0.689	0.697	0.683	0.710
500	0.695	0.705	0.698	0.729
600	0.700	0.713	0.715	0.738
700	**0.707**	0.726	**0.724**	0.750
800	0.701	**0.727**	0.719	**0.753**
900	0.700	0.721	0.712	0.750
1.000	0.698	0.721	0.703	0.746
1.100	0.693	0.719	0.699	0.745
1.200	0.674	0.720	0.691	0.745
Average	0692	0710	0698	0729
Maximum	0.707	0.727	0.724	0.753
Minimum	0.674	0.676	0.663	0.663

Highest values in each column are in bold.

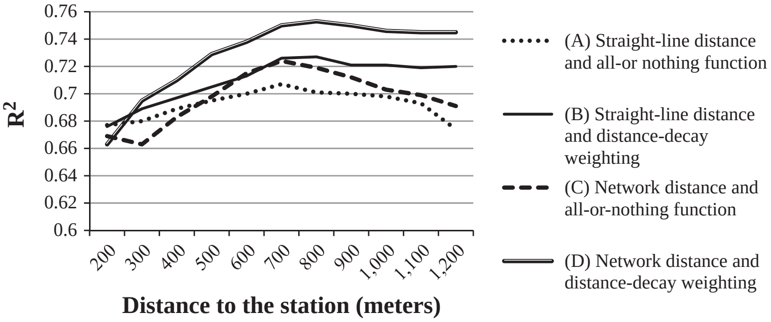

Fig. 4. Sensitivity of the model (determination coefficients R^2) to changes in the procedures for delimiting service areas, distance-decay functions and distance thresholds.

Table 6
Sensitivity of the model: differences between columns in Table 5.

Distance band	B–A		D–C		D–A		C–A		D–B	
	Abs.	%	Abs.	%	Abs.	%	Abs.	%	Abs.	%
200	−0.001	−0.15	−0.006	−0.90	−0.014	−2.11	−0.008	−1.20	−0.013	−1.96
300	0.009	1.31	0.032	4.60	0.015	2.16	−0.017	−2.56	0.006	0.86
400	0.008	1.15	0.027	3.80	0.021	2.96	−0.006	−0.88	0.013	1.83
500	0.010	1.42	0.031	4.25	0.034	4.66	0.003	0.43	0.024	3.29
600	0.013	1.82	0.023	3.12	0.038	5.15	0.015	2.10	0.025	3.39
700	0.019	2.62	0.026	3.47	0.043	5.73	0.017	2.35	0.024	3.20
800	0.026	3.58	0.034	4.52	0.052	6.91	0.018	2.50	0.026	3.45
900	0.021	2.91	0.038	5.07	0.050	6.67	0.012	1.69	0.029	3.87
1000	0.023	3.19	0.043	5.76	0.048	6.43	0.005	0.71	0.025	3.35
1100	0.026	3.62	0.046	6.17	0.052	6.98	0.006	0.86	0.026	3.49
1200	0.046	6.39	0.054	7.25	0.071	9.53	0.017	2.46	0.025	3.36
Average	0.018	2.54	0.031	4.25	0.037	5.08	0.006	0.86	0.019	2.61
Maximum	0.020	2.75	0.029	3.85	0.046	6.11	0.017	2.35	0.026	3.45
Minimum	0.002	0.30	0.000	0.00	−0.011	−1.66	−0.011	−1.66	−0.013	−1.96

A. Straight line distance and all-or nothing function.
B. Straight line distance and distance-decay method.
C. Network distance and all-or-nothing function.
D. Network distance and distance-decay method.

bands distorts the results when the all-or-nothing function is used, but does not produce a significant decrease in the explanatory power of the model if the distance-decay weighting procedure is used (since very low weights are given to people living in the bands farthest from the service area).

The way the service area is calculated (straight-line versus network distances) does not seem to have a decisive influence on the results when the all-or-nothing function is used (0.9% on average and 2.5% on the 800 m threshold, i.e., the differences between columns C and A in Table 6), but seems to be more influential using distance-decay weighting (2.6% on average and 3.4% on the 800 m threshold, i.e., the differences between columns D and B). The weak increase when the all-or-nothing function is used may be due to the fact that in a very dense network (like the Madrid Metro) the delimitation of the service areas depends much more on the limits of the Thiessen polygons (competition between stations) than on the method used to calculate the distance to the station (this only influences peripheral stations, which have no external competition). The greater improvement in the second case can be explained by the fact that more bands in the same service area (and a more detailed treatment of the distance decay) are to be expected with the network distance procedure, since Euclidean

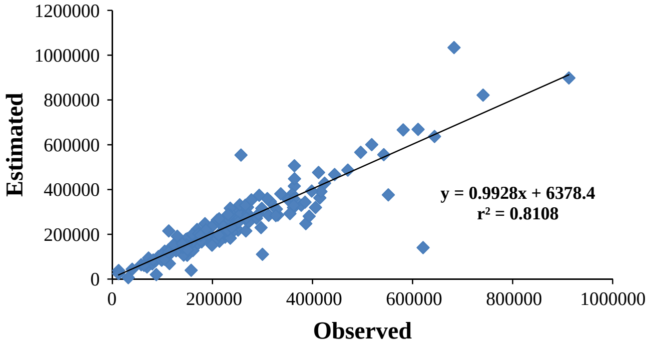

Fig. 5. Correlation between observed and expected passengers: four-step model.

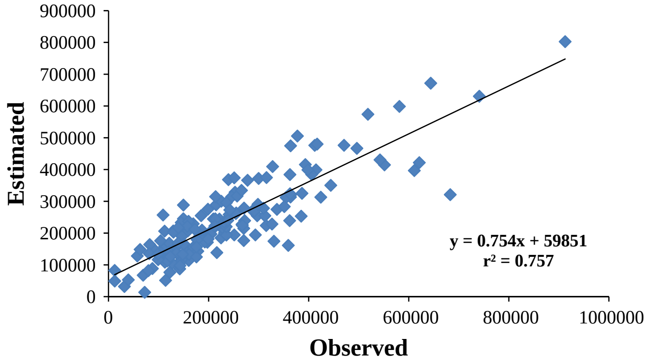

Fig. 6. Correlation between observed and expected passengers: distance-decay weighted regression.

distances systematically overestimate the size of the area covered by the bands.

However, it has still to be determined whether the proposed model has enough explanatory capacity or whether it is far off that of the four-step model. To test this point, comparisons are shown between the numbers of boardings observed (November 2004 boardings) and estimated at each station, using both the proposed direct estimation model (Fig. 6) and the four-step model (Fig. 5). The estimates for the four-step model are taken from modeling conducted by the Madrid Transport Authority (Consorcio Regional de Transportes). The correlation coefficients between observed and estimated boardings for the four-step model and for the direct model are 0.811 and 0.753, respectively. This means that the explanatory power of the simple direct model is relatively high in comparison to that of the more sophisticated four-step model.

5. Final remarks

Direct models estimate ridership as a function of the characteristics of the stations and of their service catchment areas. As an alternative to the complex and costly four-step model, they respond to a real, defined need that transport planning managers have when faced with situations requiring immediate, reliable results at reasonable cost.

Direct estimation models do this by giving an in-depth local analysis of the station service area. This allows a very detailed consideration of the station environments, working with ‘fine grain’ resolution, which is needed to analyze pedestrian access to stations. In addition, this type of model allows the testing of hypotheses on what factors influence boardings at station level. This study found nine highly significant variables for explaining ridership in the Madrid Metro, and seven clearly non-significant factors. One weakness of the model is that as it is based on pedestrian access to the stations and does not have a regional perspective, it cannot be used to carry out precise estimations for stations where access is not mainly pedestrian, as in the case of intermodal stations.

Multiple regression models consider variables calculated for station service areas and not for transport zones (which is what four-step models work with), which involves carrying out information disaggregation operations. In this context, a series of improvements has to be taken into account for calculating the variables entered in the model. The model presented considers the need to evaluate the distance threshold used (instead of using a fixed distance threshold by assimilation from other studies), the distance calculation procedure (network distance versus straight line distance) and, above all, the addition of distance-decay functions (weighting the variables accordingly).

Previous ridership models do not consider that the use of public transport tends to decrease with the distance from stations/stops within the service areas. In this paper, two distance-decay functions have been calibrated, using data from a mobility survey in order to weight service area variables. Analyses carried out show that weighting the variables according to distance-decay functions provides systematically better results. In addition, the distance-decay weighted regression model is more useful for evaluating land-use policies than other direct ridership models at station level. The paper demonstrates that the model is able to reflect the fact that concentrating housing and employment near the station will have more impact on boardings than if such developments were located farther away within the station catchment area. In contrast, previous direct ridership models at station level estimate the same increase in journeys irrespective of the band in which new developments are to be located. Instead of working with population and employment densities, our particular model considers population groups and employment types as predictors (which implies considering densities indirectly) in order to attain higher explanatory power. However, these predictors can be replaced by population and employment density variables, according to the purpose of the study. The most important contribution of this paper is not the particular model obtained, but the methodological improvements proposed.

At the same time, evidence in the paper suggests that the choice of critical distance thresholds for delimiting service areas is important in terms of demand. Instead of using a hard standard, the paper proposes choosing the distance threshold based on the correlation existing between the independent variables and ridership, so that this critical threshold can be justified from the point of view of demand. In fact, it has been verified that the use of this method significantly improves the results. The way the service area is calculated (straight-line or network distances) does not seem to have a decisive influence on the results when the all-or-nothing function is used, but seems to be more influential when distance-decay weighting is used.

In general terms direct ridership models should not be seen as substitutes for four-step models, as their application is more limited. It is more a question of having a tool available for use in situations which require a rapid response, or in cities with limited budgets. One of the most important advantages of these models is the notable cost savings in terms of money, technical personnel and the time taken to obtain results, since they do not need to be fed by a mobility survey but are obtained simply by calculating the variables in the service areas. Such models also provide complementary information to the four-step model, as they allow the influence of urban planning policies on the use of public transport to be evaluated. Direct estimation models considerably reduce complexity without jeopardizing precision.

Acknowledgements

We gratefully acknowledge the financial support from the CRTM (*Consorcio Regional de Transportes de Madrid*) and from the *Ministerio de Ciencia e Innovación* (TRA2008-06682) and also wish to thank José Antonio Cascales, Maite Antón and the two anonymous referees for their constructive comments.

References

Bernick, M., Cervero, R., 1997. Transit Villages for the 21st Century. New York, McGraw-Hill.

Bertolini, L., 1999. Spatial development patterns and public transport: the application of an analytical model in the Netherlands. Planning Practice and Research 14 (2), 199–210.

Cervero, R., 1994. Rail-oriented office development in California: how successful? Transportation Quarterly 48 (1), 33–44.

Cervero, R., 2002. Built environments and mode choice: toward a normative framework. Transportation Research Part D 7, 265–284.

Cervero, R., 2004. Transit Oriented Development in America: Contemporary Practices, Impacts, and Policy Directions. International Planning Symposium on Incentives, Regulations, and Plans – the Role of States and Nation-States in Smart Growth Planning. University of Maryland.

Cervero, R., 2006. Alternative approaches to modelling the travel-demand impacts of smart growth. Journal of the American Planning Association 72 (3), 285–295.

Cervero, R., Duncan, M., 2002. Residential Self Selection and Rail Commuting: a Nested Logit Analysis. Working Paper 604, University of California Transportation Center, Berkeley.

Cervero, R., Kockelman, K., 1997. Travel demand and the 3Ds: density, diversity, and design. Transportation Research Part D 2, 199–219.

Chakraborty, J., Armstrong, M., 1997. Exploring the use of buffer analysis for the identification of impacted areas in environmental equity assessment. Cartography and Geographic Information Systems 24 (3), 145–157.

Chu, X., 2004. Ridership Models at the Stop Level. National Center of Transit Research, University of South Florida.

Clark, W.A.V., Hosking, P.L., 1986. Statistical Methods for Geographers. Wiley, New York.

Cristaldi, F., 2005. Commuting and gender in Italy: a methodological issue. The Professional Geographer 57 (2), 268–284.

Curtis, C., Renne, J.L., Bertolini, L. (Eds.), 2009. Transit Oriented Development: Making it Happen. Ashgate, Aldershot.

El-Geneidy, A., Tétreault, P., Surprenant-Legault, J., 2010. Pedestrian access to transit: identifying redundancies and gaps using variable service area analysis. Paper Presented at the 89th, Transportation Research Board Annual Meeting, Washington, DC, USA.

Filion, P., 2001. Suburban mixed-use centres and urban dispersion: what difference do they make? Environment and Planning A 33, 141–160.

Gomez-Ibanez, J.A., 1996. Big city transit ridership, deficits, and politics: avoiding reality in Boston. Journal of the American Planning Association 62 (1), 30–50.

Giuliano, G., 2003. Travel, location and race/ethnicity. Transportation Research A, Policy and Practice 37 (4), 351–372.

Gutiérrez, J., García-Palomares, J.C., 2008. Distance measure impacts of public transport service areas. Environment and Planning B – Planning and Design 35, 480–503.

Hansen, W., 1959. How accessibility shapes land use. Journal of the American Institute of Planners 25, 73–76.

Horner, M.W., Murray, A.T., 2004. Spatial representation and scale impacts in transit service assessment. Environment and Planning B – Planning and Design 31, 785–797.

Hsiao, S., Lu, J., Sterling, J., Weatherford, M., 1997. Use of geographic information systems for analysis of transit pedestrian access. Transportation Research Record 1604, 50–59.

Jin, X., 2005. Impacts of Accessibility, Connectivity and Mode Captivity on Transit Choice. US Department of Transportation, Federal Transit Administration.

Keijer, M.J.N., Rietveld, R., 2000. How do people get to the railway station? The Dutch experience. Transportation Planning and Technology 23 (3), 215–235.

Kitamura, R., 1989. A causal analysis of car ownership and transit use. Transportation 16, 155–173.

Kuby, M., Barranda, A., Upchurch, C., 2004. Factors influencing light-rail station boardings in the United States. Transportation Research A 38 (3), 223–247.

Levinson, H.S., Brown-West, O., 1984. Estimating bus ridership. Transportation Research Record 965, 8–12.

Loutzenheiser, D.R., 1997. Pedestrian access to transit. Model to walk trips and their design and urban form determinants around Bay Area Rapid Transit Stations. Transportation Research Record 1604, 40–49.

Marshall, N., Grady, B., 2006. Sketch Transit Modeling Based on 2000 Census Data. TRB 2006 Annual Meeting CD-ROM.

Miller, H., 1999. Potential contributions of spatial analysis to geographic information systems for transportation (GIS-T). Geographical Analysis 31, 373–399.

Miller, H., Shaw, S., 2001. Geographic Information Systems for Transportation: Principles and Applications. Oxford University Press, New York.

Murray, A.T., 2001. Strategic analysis of public transport coverage. Socio-Economic Planning Sciences 35, 175–188.

Murray, A.T., Davis, R., Stimson, R.J., Ferreira, L., 1998. Public transport access. Transportation Research D 3 (5), 319–328.

Nyerges, T., 1995. Geographic information system support for urban/regional transport analysis. In: Hanson, S. (Ed.), The Geography of Urban Transportation. Guilford Press, New York, pp. 240–268.

McNally, M.G., 2000. The Four Step Model. Institute for Transportation Studies, University of California, Irvine.

ÓNeill, W.A., Ramsey, R.D., Chou, J., 1992. Analysis of transit service areas using geographic information systems. Transportation Research Record 1364, 131–138.

O'Sullivan, S., Morrall, J., 1996. Walking distances to and from light-rail transit stations. Transportation Research Record 1538, 131–138.

Parsons Brinckerhoff, 1996. Transit and Urban Form, TCRP Report 16, vol. 1. Transportation Research Board, National Research Council, Washington, DC.

Peng, Z., Dueker, K., 1995. Spatial data integration in route-level transit demand modeling. Journal of the Urban and Regional Information Systems Association 7, 26–37.

Pushkarev, B.S., Zupan, J.M., 1982. Where transit works: urban densities for public transportation. In: Levinson, H.S., Weant, R.A. (Eds.), Urban Transportation: Perspectives and Prospects. Eno Foundation, Westport, CT.

Seskin, S., Cervero, R., 1996. Transit and Urban Form. Federal Transit Administration, Washington DC.

Taylor, B.D., Fink, F.N.Y., 2003. The Factors Influencing Transit Ridership: A Review and Analysis of the Ridership Literature. UCLA, Los Angeles Transit Administration.

Untermann, R., 1984. Accommodating the Pedestrian: Adapting Neighborhoods for Walking and Bicycling. Van Nostrand Reinhold, New York.

Upchurch, C., Kuby, M., Zoldak, M., Barranda, A., 2004. Using GIS to generate mutually exclusive service areas linking travel on and off a network. Journal of Transport Geography 12 (1), 23–33.

Wachs, M., 1989. US transit subsidy policy: in need of reform. Science 244, 1545–1549.

Walters, G., Cervero, R., 2003. Forecasting Transit Demand in a Fast Growing Corridor: The Direct-Ridership Model Approach. Technical Memorandum prepared for the Bay Area Rapid Transit District. Fehr and Peers, Lafayette, CA.

Zhao, F., Chow, L.F., Li, M.T., Gan, A., Ubaka, I., 2003. Forecasting transit walk accessibility: a regression model alternative to the buffer method. Transportation Research Board Annual Meeting.

Zhu, X., Lee, C., 2008. Walkability and safety around elementary schools: economic and ethnic disparities. American Journal of Preventive Medicine 34, 282–290.