A logical approach for geo-coding and minimizing noise in the spatial data of public bus services in urban areas

Y. E. Hawas[1,2], M. B. Khan[2], N. Basu[2] & K. Ahmed[1,2]
[1]*Civil and Environmental Eng Department, United Arab Emirates University, Al Ain, UAE*
[2]*Roadway, Transportation and Traffic Safety Research Center, United Arab Emirates University, Al Ain, UAE*

Abstract

A more efficient public transit system means better accessibility and higher opportunities for better economy. To enhance the operation or to provide a better transit service, it is essential to draw a clear picture of the baseline conditions to identify the pros and cons of the existing service. Developing a quantitative spatial database of the service characteristics is deemed necessary for detailed analyses of existing services. Spatial database may help in understanding the geographical extent of the services and identifying the areas that may need special considerations to enhance the services. In this paper, a logical approach is presented to geo-code the spatial data of the public bus services; namely, the stoppage and routes maps (layers). The used geo-data data to develop the bus stops and routes' layers have significant noises. The adopted methodology to filter such noise in data and to develop the spatial database is presented in details. Analyses and conclusions were presented using the presented GIS spatial data techniques.
Keywords: geo-coding, noisy data, public transport.

1 Introduction

A public transit system is an essential part of urban life. It provides an important option that can cut through the congestion and provide access to the job market. Public transportation services provide a link to connect suburban job facilities as well as the job market in the urban core with employees. Thus it has become an


WIT Transactions on The Built Environment, Vol 116, © 2011 WIT Press
www.witpress.com, ISSN 1743-3509 (on-line)
doi:10.2495/UT110231

important economic indicator of a country and is required for meeting rapidly growing mass mobility needs. To avail such efficient service, it is necessary to have a good qualitative and geo-spatial database of the transit system itself [1]. A strong and informative database is one of the important requirements to operate any public transit service. For a new service, it is also required to establish the baseline condition of the performance.

The Geographical Information System (GIS) has already been an established technique to analyze public bus service performances. A spatial database can enhance the information related to the public transit system and helps us to understand how the overall system is performing. It is well-known that the geographic setting within which a transit system operates can strongly affect its performance and effectiveness. The characteristics of the local population, transportation network, and commuting patterns largely determine passenger demand as well as operational scale for the public transit system [2]. Therefore, it is necessary to examine the performance of public transit systems from both the managerial and geographic perspectives. The ability of GIS to integrate digital maps and spatial analytical methods has made it a powerful tool for transportation planning [3]. El-Shair [4] used GIS and remote sensing technique to identify the land-use type and share around the bus stoppage location of Birkenhead, Auckland. Meyer and Saeasua [5] showed that GIS can provide strong management decision support capability as well as an analysis tool for the authority. Many other researchers showed the challenges and prospects of using GIS techniques for transportation [6–13].

The success of the GIS tools or methods for transit systems planning or enhancing performance depends to a great extent on the accuracy of the database. Prior to the application of the analytical methods to evaluate or assess the transit system effectiveness, it is necessary to build an accurate and informative spatial database first. Basic elements of such database are Stoppage [point] and Route [line] layers.

Geo coding becomes easy and straightforward if the points and lines are well defined and can be identified. In the case that the transit system is immature or if the service is relatively new, the transit planner [being unsure about optimal bus stop locations or even route alignment, with no actual demand data exists] specifies the bus stops' locations tentatively [say within a proximity of some point] with the exact bus stop [in each trip] to be determined by the bus operator based on the actual loading/off loading demands in the field.

In the case that such bus stops are not identifiable or variant from one trip to another, a series of methods and procedures may need to be consecutively applied to geo-code the various data points representing the stops. Essentially, special methods are needed to handle problems associated with significant levels of noise in the data or discrepancy due to variations caused by inaccurate bus stop locations.

A number of studies have been done on issues of data noises and redundant data set. A new method for detecting the redundant data by plotting them on a virtual image plane and collecting the data per pixel is presented in Swadzba et al. [14]. A digital filter for noise reduction is developed by May and Fritsch

[15]. It selects between local variances obtained from adjacent pixels in the same frame and adjacent pixels in the same field. A new geospatial modelling system is adopted in by Rahmes et al. [16]. It includes a geospatial model database and a processor for noise filtering operation on elevation data associated with respective location points.

In this paper, a logical procedure is adopted to geo-code the bus stops layer. The procedure is applied on a dataset characterized by significant levels of noise and variations. The idea is to filter and consolidate the neighboring data points [within a close proximity of a bus stop location] collected during many bus trips, into a unified manageable route bus stop dataset. Following the geo-coding of the bus stops, the route line [layer] is developed using the final set of stops on each route. Analyses were carried out by incorporating the attributes' data of the stoppage and route layers. The procedures presented in this paper involved using TransCAD 4.5, ArcGIS 9.2 and Microsoft Office Excel.

The methodologies presented herein were actually developed as parts of a broader study for evaluating the public bus services in Abu Dhabi and Al Ain cities of the United Arab Emirates.

2 Data collection

There are 8 urban bus routes operating in Al Ain city and 12 routes in Abu Dhabi. Data were collected on all routes, for a week; five weekdays and two weekends (three peak periods daily). Table 1 shows the total number of conducted surveys.

Table 1: Overall samples of logs' survey.

City	Number of Routes	Overall number of collected log surveys (7days*3 peak periods* Number of Routes*2 direction)
Abu Dhabi	12	504
Al Ain	8	336
Total	20	840

A single route (# 930) in Al Ain city is taken here as an example to discuss the methodology and analyses. The majority of the challenges were faced in Al Ain city; the bus service was very recently introduced, the bus stops are not well identified. The data collection on this route included the so-called "Route Log Survey" comprising counting the passengers boarding and alighting at each bus stop, and time duration for boarding/un-boarding at each of these stops. Accompanied with the passenger counting, the GPS coordinates of each bus stop is also recorded.

Several challenges were faced before starting the survey, throughout field data collection, and afterwards during the analysis phase of the project. The identification of bus stoppages was among the great challenges for the surveyors. In Al Ain city, as the bus service has been established very recently, the bus stops are not defined properly in many places.

To resolve this issue, each surveying team (comprising two individuals) was given a GPS device while travelling the routes on the bus to collect the stoppage location (GPS coordinate data). A bus stop is recorded when the bus stops to board/ alight passengers, even if there is no post on the highway. As the survey has been done for 21 times (7 days *3 peak periods) for each route and each direction, it is assumed that all probable locations (Latitude/Longitude) of stoppages along each route would be gathered by the surveying team. That is, there is no chance that certain bus stops will not be recorded or skipped because the route is surveyed repeatedly (21 times).

Keeping in mind the variability of the bus stop locations [as indicated earlier], with no well-identifiable bus stop posts or signs, surveying each route direction *21 times [7 days a week * 3 peak hours]* ultimately results in a cluster of points for each bus stop. That is, the same bus stop may be represented by several data points, representing the route log readings during the various days and peak hours. Each data point of the same bus stop may as well have its GPS coordinates. This paper shall briefly describe the methodology adopted to identify, cluster and group the individual bus stop data points.

3 Methodology

Following below are the adopted steps to finalize the stoppage layer for a single route:

3.1 Selecting the projection system and converting the latitude and longitude of the survey data according to the selected projection system

First, a common *projection* system has been selected for the GPS devices. As the survey has been done in two cities (Abu Dhabi and Al Ain) and by a number of survey teams, it was very important to use a common *projection* to reduce error margin while plotting the data on the software. The GCS_WGS_1984 (Datum: D_WGS_1984; Prime Meridian: Greenwich; Angular Unit: Degree) projection which is being used by the Department of Transport (DOT) was selected.

The existing road layer map of Al Ain city was previously developed with the projection system- GCS_Nahrwan_1967_UTM Zone_40N (Projection: Transverse Mercator). The road layer was converted to D_WGS_1984; Prime Meridian: Greenwich; Angular Unit: Degree projection system. This step was not done for Abu Dhabi as the projection system used for the Abu Dhabi road layer was already D_WGS_1984; Prime Meridian: Greenwich; Angular Unit: Degree.

3.2 Preparing the database file for all latitude and longitude data of all peak period (7 days * 3 peak periods)

As previously indicated, 21 log files were collected for each route direction. Each trip (one direction of one route) has been treated separately. Each log file gives the location of the stoppages. The locations of the bus stops are not same among all the log files.

Stoppages are defined as the location at which the bus stops to board/disembark passengers, even if there is no post on the highway. In one peak hour, there might be 'X' number of stoppage points. For the same route and direction, there might be 'Y' number of stoppages in another peak period. The first task was to prepare a point layer consisting of all probable stoppage location as well as the associated attribute data of all stoppages.

While doing this task, two issues were raised. First issue had to do with devising a simple technique to consolidate all the survey points in a single database. The second issue was how to merge all the passengers' loading/alighting data with these stoppage points. In addressing these issues, a matrix was created in Microsoft Excel. The matrix has the first column as the bus stop ID (consecutively incremented). Another column included a unique ID [representing the route number, the day of the week and the peak hour of the survey]. The passengers loading/alighting on a specific day and a specific peak hour are inserted in separate columns. By consolidating all the collected survey logs of route 930 (one direction), it was found that such file contains a total of **693** stoppage points.

3.3 Plotting the lat/long data of the stoppages

Stoppage points (by Latitude /Longitude) were plotted on the city's road layer. For each route, two files were created; one for each route direction. A sample map showing the initial plotted stoppage locations is shown in figure 1.

As can be seen, not all points are on the designated route line. Sources of such latitude/longitude errors could be errors in GPS readings (mostly due to the GPS satellite weak coverage problems inside the bus).

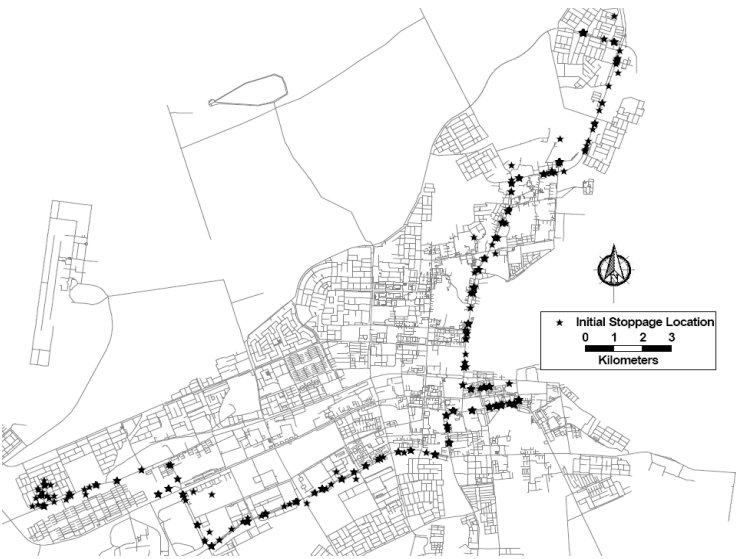

Figure 1: Location of all stoppage points (one direction) of route 930.

The next task was to make sure that the data is error free as much as possible, and hence finalizing the stoppage locations.

3.4 Preparation of initial stoppage layer, creating buffers, point data clustering and finalizing the stoppage layer

As mentioned above, route 930 stoppage layer contained a total of 693 points. A GIS-based point layer was plotted using the latitude and longitude data. The point layer was projected on the road layer.

For the 693 points we introduced only 21 ID's. For instance, an ID of 1 is used to label all stoppage points on Sunday AM, an ID of 2 is used to label all stoppage points on Sunday PM, etc. This means that same ID is used to label all data points collected at same day and same peak hour. That is, the ID's are repeated. For a stoppage at location A on Sunday AM an ID of 1 is given. The same ID of 1 is used to label another stoppage point at Location B (on Sunday AM peak hour).

The clustering of points was done using three conditions. First, each cluster should not contain more than one point of same ID. The used buffer is 100 meters. This applies then to the second condition; the distance between any two points in the same buffer should not exceed the 100 meters. Third, the maximum number of points in the same buffer should not exceed 21 points with various ID's.

To apply these conditions, initially a buffer of 100 meters radius was created around the points in the *initial stoppage location layer*. This initial stoppage layer included only the bus stops obtained during the one day and peak hour that entailed the maximum number of stops (whatever that day is and whatever peak hour is). For route 930, the initial stoppage layer contained 48 points, representing the stops [on the day and peak hour entailing the maximum number of stops]. Such initial points can be regarded as the base map *seeds* of the bus stops.

Each of the 48 initially created buffers was carefully investigated against the first condition [the ID is not repeated within the same buffer]. Also, the total number of points in each buffer was counted. If the ID is found more than once in the same buffer, it is assumed due to data error. The points [of the same ID] are identified, one is kept in the buffer, while the others are shifted to either following buffer cluster or preceding one. The decision of what points to remain and what to shift is merely done based on the location of these points within the buffer.

In case of outlier points lying outside buffer zones, such outliers are shifted inside the buffer zone based on the distribution of ID of the points inside and outside the buffer.

Some clusters of points were found without having any of the initial *seed* points [initial stoppage location point]. This was considered acceptable due to the fact that it is probable that the bus might have stopped at such cluster location in days and peak hours different from those considered for the initial stoppage layer. In such case, the initial seed layer was modified by adding one seed point at the same identified cluster location on the road line.

Among the challenges in this task was also identifying the first and last stoppage locations. The loading and alighting data were used for such purpose. The initial stoppage was identified as the point at which the passenger alighting is zero [no alighting data]. The last stoppage was identified as the one with no passenger loading data.

This task was concluded for route 930 with a total of 66 stoppage locations in one direction. Figure 2 shows the finalized stoppage locations for one direction of route 930.

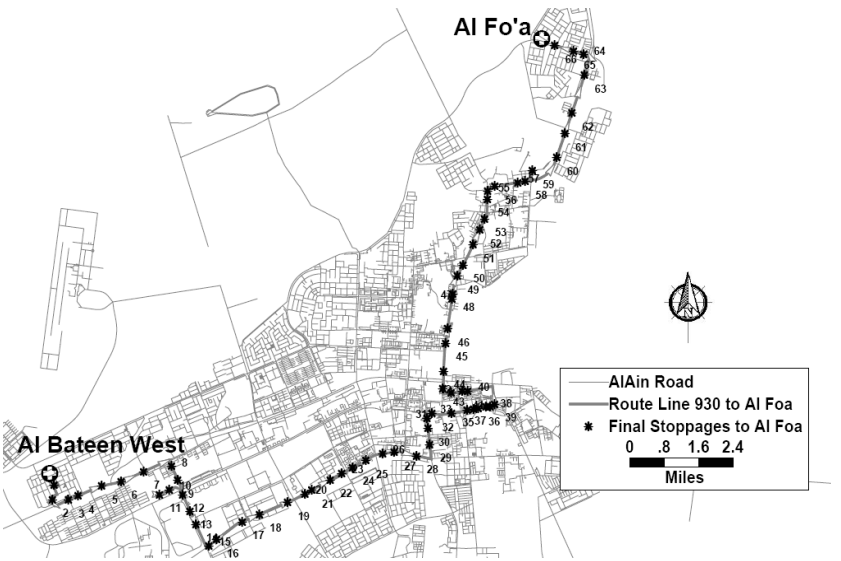

Figure 2: Final stoppage location of route 930.

3.5 Performing overlaying function to integrate the passenger loading/alighting data

The next step was to integrate all the passenger loading/alighting data with the finalized stoppage locations. The *overlay* function of the GIS environment was used to overlay the attribute data of the stoppage points to the corresponding final stoppage location. Finally, the route *line* was drawn based on the final stoppage locations.

4 Analyses

The collected loading/alighting data combined with the finalized GIS route layers were used to estimate some indicators of the system efficiency. Among these indicators were the *Bus Loading Efficiency Factor* and the *Route Utilization Indices*. The estimation was done with Microsoft Excel and a digital map for better understanding of the spatial content. Table 2 shows the basic attributes of route 930. It is to be noted that the attributes of route length,

Table 2: Route profile of route 930.

Route Number	Direction	Route length (km)	Bus Seating Capacity	Average Travel Time	No. of Stoppages
930	Inbound	93	36	1.60	66
	Outbound				60

average travel time, and number of stoppages in each direction are extracted from either the surveys or the described GIS process to finalize the stoppage layer for each route direction.

Table 3 shows the estimated utilization indices for route 930, and Table 4 shows the bus loading efficiency factors. Three indicators of route utilizations are estimated as shown in Table 4; route utilization per stop, per km of distance travelled and per hour of service. The bus loading efficiency on the other hand represents the number of on-board passengers divided by the seating capacity, at any bus stop. The maximum bus loading efficiency is the maximum number of on-board passengers at any bus stop divided by the bus seating capacity as shown in Table 4.

Table 3: Route utilization indices for route 930.

Route Number	Direction	A	B	C	Weekday & Peak Period
930	Inbound	2.33	3.31	96.25	Friday PM
	Outbound	2.15	2.77	80.63	Thursday PM

A: Route utilization per stop= Total number of boarding-in passengers/ number of stops
B: Route utilization per km =Total number of boarding-in passengers/ length of route in one direction
C: Route utilization per hour =Total number of boarding-in passengers/ travel time

Table 4: Maximum bus loading efficiency factors for route 930.

Route no	Direction	Maximum Number of Passengers On-Board	D	Weekday & Peak Period
930	Inbound	100	2.78	Friday PM
	Outbound	109	3.03	Friday MD

D: Max Bus Loading Efficiency Factor = Max. Number of Passenger On-Board/ Bus Seating Capacity.

Figures 3 and 4 illustrate the bus loading efficiency factors at each bus stop along the inbound and outbound directions of route 930, in a typical weekday. Using such an indicator within a GIS environment is extremely helpful in identifying locations or stops of high bus loading efficiencies. This will help

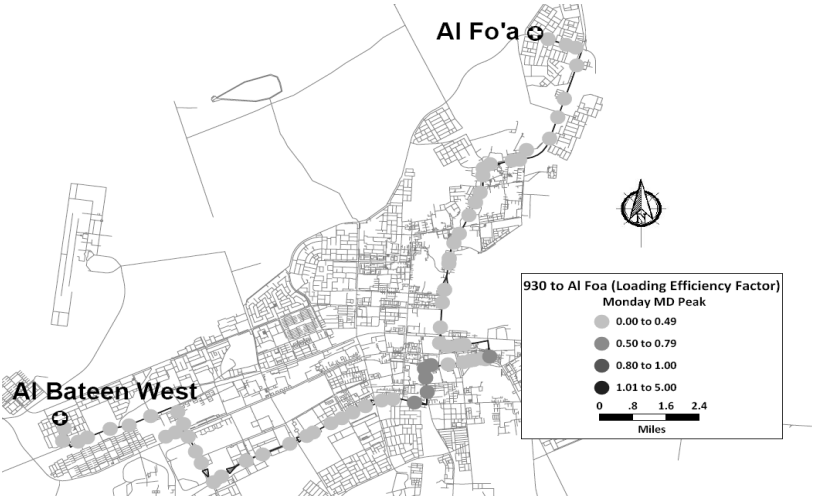

Figure 3: Route 930 (inbound) bus loading efficiency factor (a typical weekday).

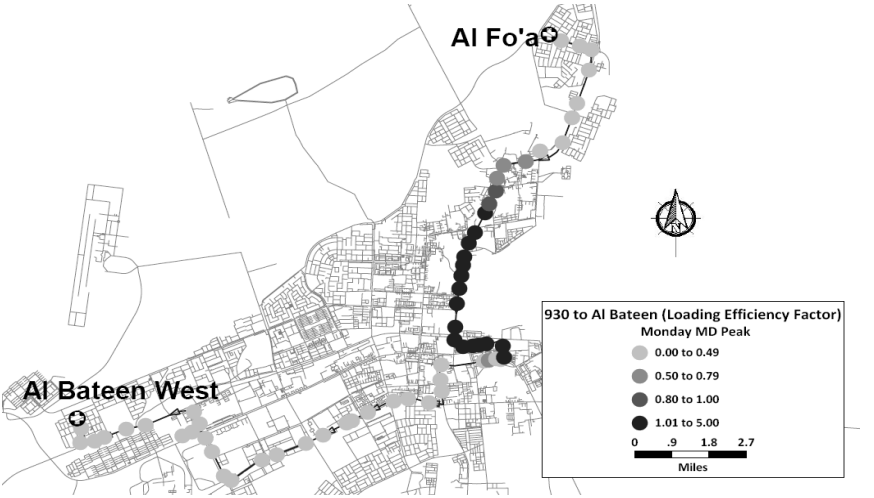

Figure 4: Route 930 (outbound) bus loading efficiency factor (a typical weekday).

In identifying the points or segments on the route where an added service could be beneficial. Higher values of efficiency factor mean more crowded bus. Such figures were helpful to conclude that the central segments of the routes are the ones that can benefit from added bus services.

Another important indicator to extract was the *bus service coverage area*. To calculate bus service coverage area, taking into consideration the local weather conditions [mostly hot weather throughout 8 months of the year], the maximum convenient walking distance of a bus user is assumed to be 300 meters. That is,

the average bus user can conveniently walk a maximum of 300 meters to reach a bus stoppage from his/her resident or work place and vice versa. Three buffering zones of 100-, 200- and 300-meter radius were used to calculate the total bus service coverage area of whole bus service in the city. A common bus stoppage point layer was created [including all stoppage locations of all routes]. This layer was then used with the buffering zones to estimate the areas of service corresponding to the 100, 200 and 300 meters buffers. Figure 5 shows the service coverage areas. The GIS capabilities were used to estimate the sum of coverage areas for each buffering radius. Table 5 summarizes the bus service coverage (in sq. km) for each buffering radius.

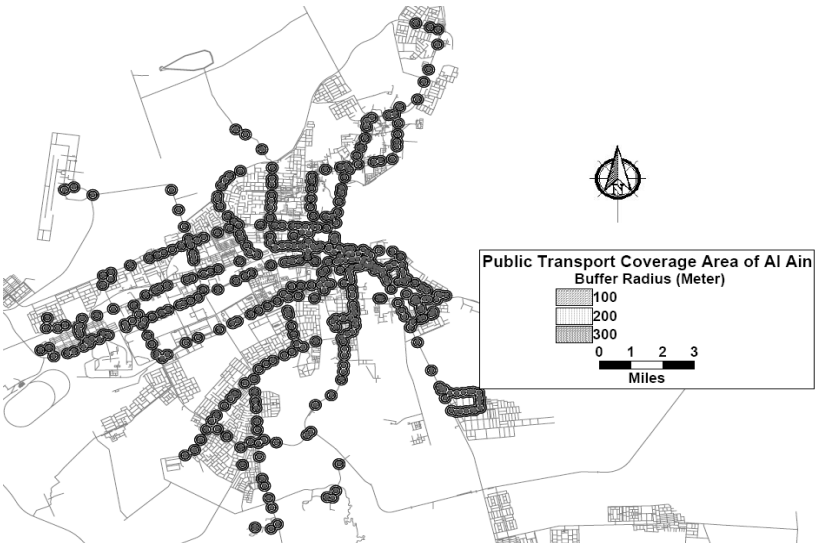

Figure 5: Bus service coverage areas for various walking distances.

Table 5: Bus service coverage areas.

Buffer Radius (meters)	Coverage Area (km^2)
100	12.02
200	42.81
300	79.68

Another useful indicator for the route performance analysis is the *passenger loading/alighting ratio* at the various bus stops. Figure 6 shows the passenger loading/alighting ratio indicator of each stoppage on all routes for one specific peak period in a typical weekday. Such indicator can be used to identify areas of heavy passenger loading or alighting. This then can be used to re-plan the routes to identify major terminal stations for instance. As can be seen in Figure 8, the central district areas and industrial area have more activities than the other areas within the city.

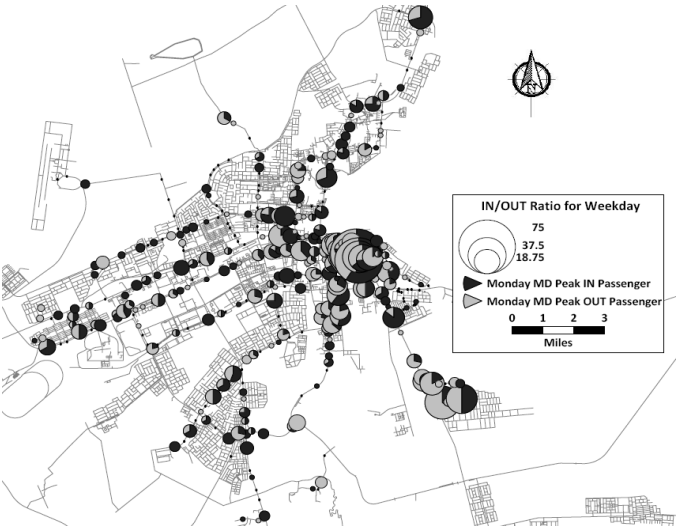

Figure 6: Loading/Alighting ratios for various bus stops in the city.

5 Concluding remarks

This paper presented a methodology and GIS-based performance indicators that can be used in any typical city with new bus services. Such tools can be used to quantify the baseline conditions that can be then used to modify or re-plan the services. The GIS-based methodology to reduce the noise in the spatial data can be generalized to develop systematic approaches for geo-coding bus stops on routes of no definite stops.

The premise of the presented methodology and the introduced GIS-based performance indicators is that they can be used [as illustrated in this paper] to assess all routes individually and provide a set of recommendations for the realigning of the various routes for more efficient performance.

Ongoing research includes the coupling of such methodology with an optimization technique to assess the relative performance of each route; identify routes with room for potential performance enhancement; and suggest changes to bus services such as frequency or headways or span of service hours to enhance the route performance indicators with minimum impact on operational cost.

References

[1] Tsamboulas, D. A., Assessing performance under regulatory evolution: A European transit system perspective, *Urban Planning and Development,* 132(4), pp. 226-234, 2006.

[2] O'Neil, W. A., Ramsey, R. D., and Chou, J., Analysis of transit service areas using geographic information systems, *Transportation Research Record*, 1364, pp. 131-138, 1992.

[3] Lao, Y. and Liu, L., Performance evaluation of bus lines with data envelopment analysis and geographic information system, *Computers, Environment and Urban System*, 33, pp. 247-255, 2009.
[4] El-Shair, I. M., GIS and remote sensing in urban transportation planning: A case study of Birkenhead, Auckland in *Conf. Rec. Map India*, 2008. Available: http://www.gisdevelopment.net/application/utility/transport/pdf/155.pdf
[5] Meyer, M. D. and Saeasua, W., Geographic information system-based transportation program management system for country transportation agency, *Transportation Research Record*, 1364, pp. 104-112, 1992.
[6] Goodchild, M.F., GIS and transportation: Status and challenges, GeoInformatica, 4(2), pp. 127-139, 2000.
[7] McComack, E. and Nyerges, T., What transportation modelling needs from GIS: A conceptual framework, *Transportation Planning and Technology*, 21, pp. 5-23, 1997.
[8] Miller, H., Potential contributions of spatial analysis to geographic information systems for transportation (GIS-T), *Geographical Analysis*, 31, pp. 373-399, 1999.
[9] Miller, H. J. and Shaw, S., Geographic information systems for transportation: Principles and applications, Oxford University: New York, 2001.
[10] Nyerges, T., Geographic information system support for urban/regional transport analysis. *Topics in The geography of urban transportation*, S. Hanson, Ed. Guilford: New York, 1995.
[11] Prastacos, P., Integrating GIS technology in urban transportation planning and modelling, *Transportation Research Record*, 1305, pp. 123-129, 1991.
[12] Thill, J., Geographic information systems for transportation in perspective, *Transportation Research Part C*, 8, pp. 3-12, 2000.
[13] Wiggins, L. et al., Application challenges for geographic information science. Implications for research, education and policy for transportation planning and management, *The Urban and Regional Information Systems Association*, 12, pp. 52-59, 2000.
[14] Swadzba, A., Vollmer, A., Hanheide, M. and Wachsmuth, S., Reducing noise and redundancy in registered range data for planar surface extraction, in the *19th Int. Conf. on Pattern Recognition*, 2008.
[15] May, W. and Fritsch, J., Structure and method for reducing spatial noise, United States Minerya System, Inc. Santa Clara, CA, 1998. Online: http://www.freepatentsonline.com/5844627.html
[16] Rahmes, M., S. Connetti, S., Yates, H. and Smith, A. O., Geospatial modelling system for performing filtering operations based upon a sum of differences of a given and neighbouring location points and related methods, United States Harris Corp. Melbourne, FL, 2009. Online: http://www.freepatentsonline.com/7487046.html2009