A Method for Inductive Estimation of Public Transport Traffic using Spatial Network Characteristics

Simon Scheider, Michael May
Fraunhofer Institut Intelligente Analyse- und Informationssysteme (IAIS),
Sankt Augustin, Germany

Abstract. An inductive method based on spatial network characteristics and geographical neighbourhood is proposed, which allows to extrapolate a sample of representative passenger counts for an existent public transport network. The most important predictors can directly be deduced from a logical model of the spatial network. Furthermore, the attractiveness of stops can be modelled by spatial densities in their geographical neighbourhood. The method is formally specified using a typed logic.

INTRODUCTION

State-of-the art techniques for transportation forecasts normally rely strongly on deductive methods. In these approaches, traffic prediction is done in a 'four step' process of urban transportation planning: trip generation, trip distribution, mode choice and route assignment (Ortuzar 2001). This means that 4 predictive models, a traffic demand model, a trip distribution model (e.g. a gravity model), a modal model (transportation type) and a route assignment have to be interlinked, and thus the error of each one sums up for the whole model. For instance the gravity model distributes trips over the network according to a theoretical distribution (Erlander 1990). It is known that such deductive models can inherit a danger of estimation bias.

Flyvbjerg et al. (Flyvbjerg 2006) showed in a study about 210 infrastructure planning projects that the inaccuracy of traffic forecasts with this approach can be immense: Rail passenger forecasts did show a strong bias, and the forecasts were found to be inaccurate by a high margin (72% of all rail traffic forecasts had a relative error greater than 66%). Nevertheless, the method of deductive transportation forecast is especially useful and designed for planning and engineering of new traffic infrastructures, that is to estimate future traffic under currently non-existent circumstances.

In this paper, we propose a different methodological task, which gives reasons for an inductive (and simpler) forecasting method: How can public transport passenger frequencies be extrapolated in the network, that means, predicted for unknown (i.e., unmeasured) parts of an existing network infrastructure under fixed socioeconomic circumstances? Predictions should be possible for unknown cities, or unknown regions. As recent results of 'space syntax'-related research indicates, the spatial network configuration is an important predictive factor for public transport traffic. Furthermore, there seems to be an additional influence of the geographical neighbourhood of a station (Chiaradia 2005). On the empirical basis of a sample of representative passenger measurements, it should therefore be possible to train a statistical model that primarily uses simple spatial network characteristics and geographical neighbourhood descriptions as predictors for passenger frequency. The authors used this method in a project to build media performance parameters for arbitrary outdoor media locations in Germany based on estimated traffic frequencies on an extensive network.

The data model of the spatial network is considered as an important part of the method. Therefore an abstract specification of this data model will be given, which can be implemented in various ways. In the next section, abstract operations and data types of the transport network are specified. Afterwards, the feature set for prediction can be expressed in a formal way. Using this feature set, several statistical models were trained and evaluated with machine learning algorithms on 781 local passenger counts in 3 german cities.

AN ABSTRACT SPATIAL NETWORK DATA MODEL

The basic data model

In the following, data types are expressed as sets by upper case symbols. Data structures and operations are formal expressions of a typed higher order logic, as in Scheider 2007. In this language, arbitrary complex data structures for geometrical and graph logical properties can be conveniently specified together. Syntax and semantics of this language are equal to the 'Basic Extended Simple Type Theory' (BESTT) approach (Farmer 2003 and Farmer 2004), so that each type stands for a set, and derived types can be constructed from atomic ones by combining them to tuple, function, set and list types, meaning their appropriate set theoretic combinations (compare table 1). Let A and B be any two types. In this formal approach, every expression 'a' has a type 'A', that means a is a member of the set denoting A, and they can be written together like this, a:A. Expressions can be formed by typed variables, constants, function abstraction (written as 'λ[variable].[expression]') and function application (written as '[function]([expression])'), so that function arguments are of the appropriate type. Some language constants with obvious meaning are used, like:

||:set[A]→Rat (:=set cardinality), ||:list[A]→Rat (:=list size), #i: $A_0\times\ldots$, $A_i,\ldots\times A_j\rightarrow A_i$ (:= extracts the i^{th} element of a tuple), [i]:list[A]→A (:=extracts the i^{th} element of a list), if: Bool×A×A→A (:=conditional expression), I:(A→Bool)→A (:= finds a unique object of type A by a predicate (definite description)), as well as well known logical and arithmetic symbols. The basic types of the basic data model are introduced in table 2 and in the following text.

Type constructors	Comments
A×B	Tuple type
A→B	Function type
set[A]	Set type
list[A]	List type
P[A]	Power set type

Table 1: Type constructors for data types

The basic data model of the spatial network is a partially embedded graph without loop edges. Let 'N' be the data type of nodes of a public transport network, so each node stands for a public transport stage. Let 'point' be a total function from 'N' to the 2-dimensional Euclidian space 'P', so that each stage has a geographical reference (a geometrical embedding). Let 'E' be the type of edges of the network as the cross product of 'N', that is E:=N×N. Linear geometries for edges of a public transport network are not necessary for the proposed method, and therefore the graph needs only a 'partial' embedding. Furthermore, edges that connect a node to itself are not allowed.

There are some additional data types. Let 'POI' be a type for 'points of interest', that is spatial point locations of attractive facilities, e.g. touristic and cultural facilities, public utilities, hotels, and so on. A subset of this type can be considered as a layer of geometrical points. Let 'point' be also a function from 'POI' to 'P'. Let 'Rat' be the type of a rational number, and 'T' be the type of a timestamp.

Data types	Comments
N	Nodes
E	Edges := N×N
T	Time stamps
Rat	Rational numbers
P	2-dim Euclidian space:= Rat×Rat
POI	Points of interest

Table 2: Basic data types

Building the connectivity network graph from raw data

A graph as a model of the public transport network infrastructure can be derived from a digital schedule of each public transport line. In analogy to Güting (1994), such a line can be modelled as a collection of paths through a line's connectivity graph. Suppose a schedule for a certain line '1' is a list of paths, each path being a connected list of edges through the network, that is a sequence of pairs of stops of the public transport line '1', with each of the two stops having a timestamp (data type T) for departure and arrival time:

(3) PATH : list[E×T×T]

(4) SCHEDULE : list[PATH]

A connectivity graph is of type G : P[E]×P[N]. Let E_1 be the set of edges of the paths of line '1', and N_1 be the set of nodes in E_1. Then the connectivity graph of line '1' is $(g_1: G) := (E_1, N_1)$.

Let L = {1,…,m} be a set of numbers being public transport line descriptors of a network. The list of connectivity graphs of these lines is given by $Glist_m$: list[G], and the list of schedules is given by $SCHlist_m$: list[SCHEDULE]. A generalised connectivity graph can then be derived by the union of all subgraphs $G_m := (E_m, N_m)$, with $E_m := \cup_{i=1}^{m} (E_i)$ and $N_m := \cup_{i=1}^{m} (N_i)$.

Network operator for schedule frequency of lines

The simplest and most important feature for this model is the frequency with which a line passes a certain edge. This frequency is a strong indicator for traffic pressure between stops. For an edge, it can easily be derived from the line's schedule by counting the number of times this edge occurs in all paths of the schedule. Additionally, the number of stops for all lines in L can be calculated by summing up the frequency for all lines:

(5) frequency : SCHEDULE×E→Rat
:=λs:SCHEDULE, e:E. $\sum_{i=0}^{|s|-1} \sum_{j=0}^{|s[i]|-1}$ if(#1(s[i][j])= e, 1,0)

(6) nStops : E→Rat
:=λe:E.$\sum_{i=0}^{m}$ frequency($SCHList_m$[i], e)

Network operator for network influence on stops

In order to use the characteristics of the logical network for a passenger prediction model, a 'transport network influence' shall be derived. This is a synthetic feature that implies how much a network node is supposed to be under traffic pressure according to the network logic.

The first step is an operation to identify central nodes. Central stops normally have many lines passing them, therefore this could be taken as an indicator. But the problem is that the absolute number of lines stopping at one stage is strongly dependent on the width of the network and the overall number of lines existent. In order to derive a network-independent relative measure, the absolute number has to be statistically normalized. This can be done by using a high quantile, e.g. $Q_{95\%}$, of line numbers for stops of a public transport network, and to classify all stops with line numbers greater than this quantile as 'central':

(7) $\text{nLinesPassing}: N \rightarrow \text{Rat} := \lambda n{:}N.\sum_{i=1}^{|L|} \text{if}(n \in N_i, 1, 0)$
(a function for calculating the number of lines passing a stage)

(8) $\text{centreness}: N \rightarrow \text{Rat} := \lambda n{:}N.\text{if}(\text{nLinesPassing}(n) > Q_{95\%}), 1, 0)$
(a function for central nodes)

The second step is to estimate traffic pressure in the transport network as a function of logical network distance from central nodes. The network distance between all nodes can be calculated by using reachability matrices. These can be derived from the adjacency matrix of the generalised connectivity graph. A matrix can be notated as 3-valued relation. Let $A{:}N \times N \rightarrow \text{Rat}$ be the adjacency matrix of G_m, so that A(i,j) =1 iff node i and node j are connected by an edge, and and 0 otherwise. Then the reachability matrix of pathlength k, with $k \in \{1,\ldots,n\}$, is given by raising A to the power of k by matrix multiplication. For the set of reachability matrices we write $\{A^1, \ldots A^n\}$. The maximum pathlength in our approach was n=6, so that $A^6(i,j) = 1$ means that nodes i and j are path connected by 6 edges. Then the distance matrix of maximum pathlength n is a matrix of the minimum pathlengths in all reachability matrices 1 to n:

(9) $D^n :N_m \times N_m \times \text{Rat}$
$:=\{(i,j,\text{MIN}_{k=1}^{n}(\text{if}(A^k(i,j)>0,k, \text{if}(i \neq j,n+1,0)))|i,j \in N_m\}$
(a logical distance matrix giving the path-distances between all nodes with maximum pathlength n)

This distance measure from central nodes can be squared and inverted in order to give a normalised 'proximity' between 0 and 1, so that unconnected nodes have proximity 0 and nodes with distance 0 have proximity 1:

(10) $\text{prox}^n :N_m \times N_m \rightarrow \text{Rat}$
$:=\lambda i{:}N_m, j{:}N_m.\text{if}(D^n(i,j)=0, 1, \text{if}(D^n(i,j)= (n+1), 0, 1/(D^n(i,j))^2))$

Then an 'external network influence' for a node is just the maximum product of centreness and proximity for all combinations with all other nodes of the network. This means that nodes are being influenced if at least one centre is close. All nodes that are not at least n edges away from a center have an influence of 0. Nodes are increasingly influenced by a center in a quadratic manner.

(11) $\text{ext\_influence}^n : N_m \rightarrow \text{Rat}$
$:=\lambda i{:}N_m.\text{MAX}_{j=0}^{|Nm|-1}\text{if}(i \neq j,\text{centreness}(\text{list}(N_m)[j])*\text{prox}^n(i,\text{list}(N_m)[j]),0)$

Spatial operator for the attractiveness of stops

Passenger traffic emerging in a network is also a result of network external factors. Because these factors often depend on the geographic neighbourhood of the network, its spatial embedding is also relevant to modelling. The geographic neighbourhood can be seen as the source or sink of traffic. In the proximity of sources and sinks, traffic pressure normally is increasing because they are common starting or end points for many individual trips.

Network external traffic sources and sinks can be modelled from several kinds of POI using their spatial densities, that is their numbers inside of a spatial radius around stops:

(12) POIDensity : $N_m \times set[POI] \times Rat \rightarrow Rat$
:= $\lambda i:N_m$, poi_layer:set[POI], radius:Rat.
|{poi ∈ poi_layer| vectorlength(point(poi)-point(i))<radius}|

Additionally, a geographic neighbourhood with high density of population is also a candidate for traffic source or sink. Therefore, population numbers were taken from a commercial dataset of statistics on postal code areas, and were aggregated to densities on community level. For reasons of space, this is expressed as a function without any specification:

(13) popDensity : $N_m \rightarrow$ Rat

A PREDICTION MODEL AND ITS QUALITY

Several predictive statistical models were fitted to the following feature dataset F. It covers part of the public bus transport network in 3 minor german cities (‘Kaiserslautern’, ‘Heidelberg’, ‘Ulm’), with ‘pass_freq_day’ being an empirically measured passenger frequency per day for several public transport lines and a subset of 781 edges in the network:

(14) F := { ((linenumber:Rat), (e:E), (pass_freq_day:Rat),
frequency($SCHlist_m$[linenumber], #1(e)),
nStops(#1(e)), $ext\_influence^6$(#1(e)),
centreness(#1(e)),
pop_Density(#1(e)),
POIDensity(#1(e), Tourism:set[POI], 2000),
…)
| linenumber ∈ L ∧ e ∈ $E_{linenumber}$ }

Distribution parameters	bus passenger frequency per day
Mean	745.61
Standard deviation	855
Minimum	0
Maximum	4774
Total number of measurements	781

Table 3: Distribution parameters of the training set.

The distribution parameters for passenger frequency per day in F are given in table 3. The average rounded passenger frequency is 746 passengers per day, and the rounded standard deviation is 855 passengers. The first model for this dataset is the following simple linear regression. It was generated using the linear regression algorithm of WEKA[1] (using the Akaike criterion for model selection, Witten (2005)):

(15) pass_freq_day
:= 2.59 * frequency($SCHlist_m$[linenumber], #1(e)) +
-0.27 * nStops(#1(e)) +
8.41 * $ext\_influence^6$(#1(e)) +
0.33 * pop_Density(#1(e)) +
-13.07 * POIDensity (#1(e), Culture:set[POI], 2000) +
26.18 * POIDensity (#1(e), Tourism:set[POI], 2000) +
-184.54

The model quality was estimated by 10-fold cross-validation and is shown in table 4. The linear regression model has a relative absolute error[2] of 59% and a correlation coefficient of 0.75.

Quality parameters	Univariate Linear Regression	Linear regression	M5' model tree	M5' model tree (without 9% outliers)
Correlation coefficient	0.69	0.75	0.84	0.91
Mean absolute error	414.36	360.25	300.05	206.55
Relative absolute error	68.19 %	59.28 %	49.38 %	39.61 %

Table 4: Quality parameters by 10-fold cross-validation.

[1] A popular machine learning toolbox.
[2] The relative absolute error is calculated by dividing the sum of absolute differences between prediction and test values by the simple predictor error, which ist the sum of absolute differences between the test values and their average.

A better quality can be reached fitting a M5' linear model tree (Wang 1997) in WEKA. Here, the relative absolute error is 49% and the correlation coefficient is 0.83 (compare table 4 and figure 1). The absolute error, 300 passengers per day, is small compared with the standard error of the variable distribution, which is 855 passengers per day.

As can be seen in figure 1, which visualises the prediction errors by residuals, the M5' model has a good fit for the majority of values, but there are several distinct outliers with very strong deviation, which obviously is hard to fit. If the 9% strongest outliers are being removed, then the M5' model quality increases very significantly, reaching a correlation coefficient of more than 0.9. An estimation bias is not indicated by these results, because the mean error (average difference between predictions and test values) is 14 passengers, which can be considered near 0.

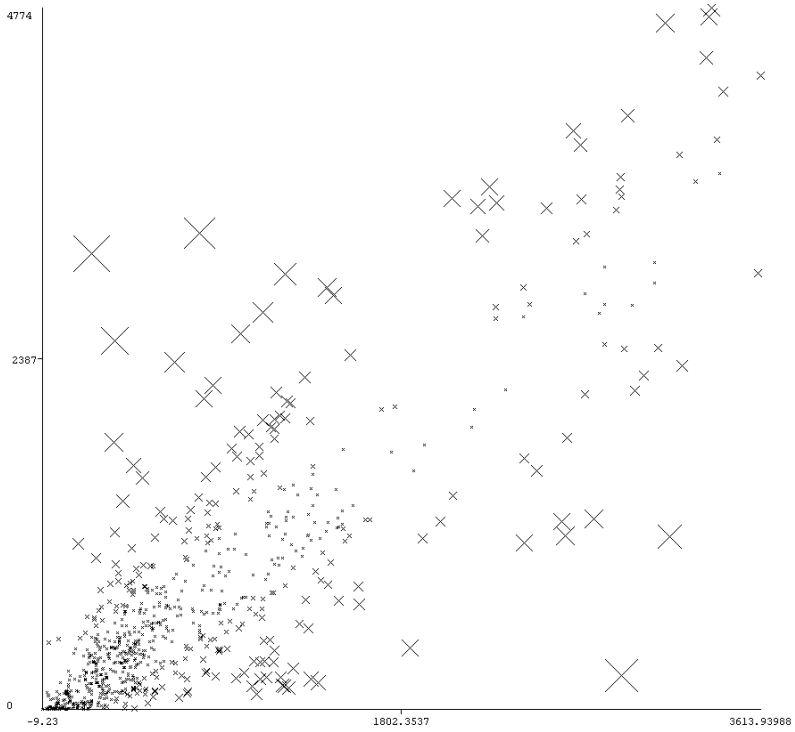

Figure 1: Prediction errors of the fitted M5' model tree. Larger crosses mean larger errors. The X axis shows predictions, the Y axis shows test values of passenger frequencies per day.

CONCLUSION

In order to extrapolate a sample of representative passenger counts for an existent public transport network infrastructure, an inductive method based on network characteristics and geographical neighbourhood was proposed. The most important predictors for this method can directly be deduced from a logical model of the spatial transport network, e.g. passage frequency of a public transport line and external influence of traffic pressure according to the network structure.

Passage frequency alone can explain up to ½ of the traffic variance of a line (compare univariate linear regression in table 4). A M5' model of all features instead can explain more than 70% of the traffic variance, with a relative absolute error less then half of the error of a simple mean prediction. Furthermore, and in contrast to many deductive forecasting methods, the estimation is not biased. Nevertheless, there are some obvious outliers, which can hardly be predicted with this simple model. It seems that these outliers exist according to currently unknown statistical influences which are still uncovered in the current feature set. Future efforts should try to incorporate them into the method. If this is not possible, perhaps the representativeness of the chosen sample of measurements should be further analysed. Furthermore, a direct comparison to deductive traffic forecasting methods should be done and discussed using a comparable dataset.

ACKNOWLEDGEMENTS

We would like to thank the German professional association for outdoor advertising ('Fachverband Außenwerbung' (FAW)) for making this study possible and for the many fruitful discussions on this topic.

BIBLIOGRAPHY

Chiaradia, A., Moreau, E., Raford, N., Configurational Exploration of Public Transport Movement Networks: A Case Study, The London Underground. 5th International Space Syntax Symposium, 2005, 541-552

Erlander, S., Stewart, N.F., 1990 The Gravity Model in Transportation Analysis – Theory and Extensions: Topics in Transportation. VSC. ISBN 9067640891, pp 226.

Farmer, W.M., A Basic Extended Simple Type Theory. SQRL Report, Vol. 14, 2003

Farmer, W.M., Mohrenschmidt, M., Simple type theory: Simple steps towards a formal specification. In 34th ASEE/IEEE Frontiers in Education Conference, Savannah, GA, F1C-6, 2004

Flyvbjerg, B., Skamris, M.K., Buhl, S.L, Inaccuracy in Traffic Forecasts. Transport Reviews, Vol. 26, No. 1, 1-24, 2006

Güting, R.H., GraphDB: Modeling and Querying Graphs in Databases. In Proceedings of the Int. Conference on Very Large Databases, 297-308, 1994

Ortuzar J., Willumsen, L.G., 2001 Modelling Transport, 3rd Edition. John Wiley. ISBN 0471861103, pp 499.

Scheider, S., A System of Abstract Specifications for Spatial Data Models. (in preparation), 2007

Wang, Y., Witten, I.H., Induction of Model Trees for Predicting Continuous Classes. Proceedings of the Poster Papers of the European Conference on Machine Learning, University of Economics, Faculty of Economics and Statistics, Prague, 1997

Witten, I.H., Eibe, F., 2005 Data Mining: Practical Machine Learning Tools and Techniques- 2nd Edition. Morgan Kaufmann. ISBN 3446215336, pp 386