ACCESSIBILITY AND URBAN DEVELOPMENT: A GRID-BASED COMPARATIVE STATISTICAL ANALYSIS OF DUTCH CITIES

Maria Teresa Borzacchiello
Department of Transportation Engineering "Luigi Tocchetti"
"Federico II" University
Via Claudio, 21,
80125 Naples
Italy
E-mail: mborzacchiello@unina.it

Peter Nijkamp
Department of Spatial Economics
Vrije Universiteit
De Boelelaan 1105
1081 HV Amsterdam
The Netherlands
E-mail: pnijkamp@feweb.vu.nl.

Eric Koomen
Department of Spatial Economics
SpinLab
Vrije Universiteit
De Boelelaan 1105
1081 HV Amsterdam
The Netherlands
E-mail: ekoomen@feweb.vu.nl

Abstract

Accessibility has become a key issue in modern urban planning. This paper aims to identify the impact of differences in spatial accessibility on the development of the built environment in cities. Using a few simple accessibility indicators, it tries to map out in a quantitative way the detailed implications of accessibility conditions for built-up areas, on the basis of a 25x25m grid cell approach. The statistical tools used are discriminant analysis and logistic regression, followed by a GIS representation of the empirical results for four Dutch cities: Amsterdam, The Hague, Rotterdam, and Utrecht.

1. Aims and Scope

It is broadly recognized that the urban land use system and the transportation system influence each other in a dynamic and complex manner; already for several decades researchers from different disciplinary backgrounds have tried to address the complex issue of accessibility (see, e.g., Blunden, 1971; Timmermans, 2003; Reggiani, 1998, de la Barra, 1989). The aim of the present contribution is to analyse land use and transportation linkages through the statistical correlation of empirical aggregate indicators stemming from both fields. The study is included in a wider framework whose aim is to obtain a data-based (in the sense that its parameters are calibrated on real world data) analytical tool (Schoemakers and van der Hoorn, 2004). Such a tool can be used to interpret, and eventually forecast, land use and transportation changes in future time periods and under different scenarios, while taking into account their mutual interactions by means of appropriate statistical parameters. These forecasts may support land use and transportation planning decisions by means of assessment with the help of indicators that are useful to understand complex land-use transportation scenarios in a clear and communicative way.

The literature about accessibility is growing in importance. This is especially the case in recent years, during which sustainable development is the 'leitmotif' of much research about land use and transportation planning. The concept of accessibility is indeed one of the best ways to integrate the mutual and complex relationships between land use and the transportation system in cities. But accessibility needs to be operationalized and, consequently, many kinds of accessibility indicators have been proposed in the literature. In this paper, the focus will be not on the design of new or more effective calculations of an accessibility indicator, but rather on stressing the relationship between accessibility and urban development. Furthermore, the dynamic and heterogeneous nature of these complex spatial phenomena calls for appropriate statistical analyses based on a spatial and temporal contextualization. And therefore, the accessibility study in this paper will be performed in a rigorous statistical way, choosing the four major Dutch cities as study areas. This choice for a comparative approach is instigated by the idea that it is interesting to analyse the same phenomenon in different urban contexts, in terms of both land use/transportation planning systems and actual urban development. Statistical analyses are performed to assess the influence of accessibility on urban development, in order to evaluate the importance of accessibility as an explanatory variable for the development of built-up areas in cities.

A quick scan of the literature on the influence of accessibility on residential and industrial location choice, brings to light that the number of solid empirical studies about the impact of accessibility on urban land use is actually rather small, while most of them are related just to residential location choice, whereas – in contrast – there are many empirical studies considering the influence of land use on the transportation system. This limited number of studies is explained by the relatively low significance of accessibility changes in urban areas in developed nations and by the empirical difficulties inherent in the estimation of accessibility, which are related to the time lag between transportation impacts and land use change (Zondag and Pieters, 2005). Furthermore, since the available empirical studies suggest that the influence of accessibility on urban residential location choice is positive, though rather modest, compared with the impact from other

demographic, social and urban factors, it may be rather hard to empirically identify this relatively small influence. Moreover, there are other methodological issues highlighted in the literature, notably the lack of the transferability of the results due to the variety of empirical applications to different regions or zones, and the ambition to include all possible explanatory variables in one comprehensive analysis (Zondag and Pieters, 2005).

This paper addresses the influence of accessibility on the development of built-up areas, by focusing on the characteristics of their spatial distribution; therefore, the results of our study are able to generate detailed localised statistical information, rather than pure statistical information. After a brief account of the accessibility concept, in the present paper some simple accessibility indicators are selected and described (essentially based on the distance to central or main facilities in the city), while next their relationship with the presence of built-up areas (as an indicator of urban development) will be investigated. The huge amount of data available from our cases has been organized in a spatial database; and an integrated data structure for subsequent analysis at different spatial scales has been set up, using a detailed spatial scale (grid cells of 25 by 25 metres), and storing all available data for different time horizons. Appropriate statistical analyses, such as discriminant analysis and logistic regression analysis, are then performed by means of suitable statistical software, using as dependent variables the development of built-up areas, and as independent variables the accessibility indicators selected, as well as some complementary land-use policy indicators.

The following methodological steps are therefore undertaken in this study:

- design of a methodology for data-based inspection of the relationships between urban land use and the transportation system;
- analysis of the statistical relationships between accessibility indicators and built-up areas for each city under consideration;
- comparison between the various case studies, in order to interpret the phenomena investigated at a wider European scale.

The paper is organized as follows. In Section 2 we will offer a brief and selective account of the accessibility concept. Then, in Section 3 the selected accessibility indicators are described, while the data structure for our urban case studies is presented in Section 4. In Section 5, the methodological rationale of our applied statistical analyses is outlined, and in Section 6 the outcomes are considered and interpreted. Finally, conclusions are offered in Section 7.

2. The Accessibility Concept

This paper focuses on the influence of accessibility on urban built-up areas. Thus, its aim is not to propose a new accessibility indicator, but instead to use a simple one in order to explore its relationships with the built environment in cities. We will start here by providing a very concise account of some prominent existing accessibility indicators, while referring for a wider review to relevant publications such as Geurs and van Wee, 2004; Geurs and Ritsema, 2001; Martellato and Nijkamp, 1999; Reggiani, 1998; Koenig, 1980.

There are many definitions of the accessibility concept, as well as a great amount of accessibility measures. Generally, there are three well-known approaches for the computation of accessibility indicators (Ettema and Timmerman, 2007; Geurs and Ritsema, 2001): the *infrastructure*-based approach mainly takes into account the performance of the transportation system; the *activity*-based approach deals with the consideration of the spatial distribution of activities, while the most recent *utility*-based approach focuses on the utility that individuals receive from accessibility to a particular destination.

From a practical point of view, in studying and computing accessibility measures, four different characteristics are often recognized (Geurs and Ritsema, 2001; Martin and Reggiani, 2007): a *transport* component related to the impedance, that is the effort necessary to reach a given destination from a given origin; a *land-use* component dealing with the attractiveness of the destination; a *time* component addressing the specific time period in which the measure concerned is observed; and an *individual* component tied to the perception and the opportunities of individuals and therefore to the relevant socio-economic system. Existing accessibility measures involve at least one of the first two components, while the last two features are present in a smaller number of indicators.

In this paper, we address two issues in particular: the specification of the relationship between the accessibility index used and the presence of the built-up areas, and the specification of the distance thresholds beyond which the influence of the accessibility of the relevant urban and infrastructure centres becomes negligible.

The first issue is related to the specification of the distance decay function; this may differ according to transport mode, purpose of trip, characteristics of the households and characteristics of the destination. Since in this study very small areas of land are regarded as origins, and since very small distances (in fact, spatial scales at less than walking distances) are taken into account, only the characteristics of the destination will be considered in order to make a distinction between the different distance decay functions.

The second issue is a common one in the specification of accessibility measures: the threshold choice is often a subjective one, and it may, of course, depend on the study aims. Usually, these thresholds are selected from statistical surveys on commuting distances or travel times. This is good practice in the case of large statistical surveys on population preferences, since the maximum distance usually forms an input into the model, based on the most accepted distance by the population. But these surveys are often not available, so that one of the tasks of our statistical analysis is the computation of the distance threshold beyond which the indicator is no longer significant, thus obtaining an objective threshold, which is, of course, strongly related to the spatial context under consideration.

A further practical issue to be dealt with is the choice of the type (zones or raster grid cells) and the width of the origin – destination area. Using zones has the advantage that it is possible to deal with a small number of origins, for which is easier to have aggregated data and official statistical surveys at various scales, but it also has some disadvantages, in particular the lack of information inside the zones. Besides, the scale is fixed, and sometimes administrative borders do not correspond with the borders of the zones considered for transportation studies. Raster representation is useful to

overcome the demarcation zone problem, but it has to deal with the choice of the scale and with data disaggregation, usually surveyed at a zonal scale (see Rietveld and Bruinsma, 1998). In our case study we are lucky enough to possess detailed geographical information that allow for a fine-scaled raster representation.

3. Selection of Accessibility Indicators

From the broad range of accessibility indicators discussed before, we select simple Euclidean distances to the main urban centres and infrastructure points and lines. These can be classified as *contour* activity-based measures (Geurs et al., 2001). The advantage of these measures is that they are rather simple to compute, although they do not take into account the perception of users of the various types of urban infrastructure nor the way the distances are perceived. Nevertheless, this kind of indicator is particularly suitable for straightforward statistical analyses, because there are not so many inputs and influences to take into account. From an operational point of view, these distances are computed here by considering as a destination point various important nodes of transportation facilities and urban centres, and as an origin each 25 by 25 m grid cell used to spatially represent the study area concerned. The transportation facilities chosen here are the regular railway stops, the Intercity stations, the highway exits (and entrances), the railway lines and the highway lines. We furthermore distinguished a number of spatial variables to represent the urban context and spatial planning regulations. These include the distance to the city centre, for which we selected the historic foundation point, the presence of natural barriers (major rivers) that divide the towns, and specific zoning regulations related to noise contours or open space preservation. In Table 1 below a concise description of the accessibility measures and other explanatory variables is provided. These variables are used in a binomial logistic regression for the dependent variable that indicates whether a grid cell is built-up (1) or not (0). It should be added that we exclude from our analyses the cells that directly refer to the highway and railway areas, since these are, by our definition, not built-up.

4. The Data Structure for Different Case Studies

The four different case studies in our analyses have a similar data organization in an ESRI geo-database: the data for the four individual Dutch cities (Amsterdam, Rotterdam, The Hague and Utrecht) are directly obtained from a series of national data sets. The input data for each case study represent:

- the area of interest, chosen with the criterion of the minimum bounding rectangle applied to the city borders;
- the railway stations, including the Intercity stations (points);
- the highway exits (points);
- the highway lines (represented by their surface area or polygon);
- the railway lines (polygon);
- the location of the city centre (point);
- the areas of restricted urbanization subjected to the buffer-zone open-space policy (polygon);

In order to obtain suitable data to perform our statistical analysis, the input data are processed uniformly and systematically, in accordance with the following flow diagram (see Figure 1). In GIScience, these schematic representations or cartographic models are common tools describe the spatial analysis process (Tomlin, 1990). It should be noted that all data refer to 2000, except for the city of Utrecht, where municipality borders considered correspond to the year 1999, thus not taking into account the subsequent annexation of a large non-urban area.

As regards the dependent variable, which describes whether a cell is built-up or not (1/0), it is derived from topographical maps of the Dutch national topographic survey.

Table 1. Accessibility measures and other explanatory variables used in the four Dutch case studies

Variable	Name	Type	Measure unit
Euclidean distance to the city centre	dis_city_centre	Continuous	km
Euclidean distance to the nearest regular train station (not Intercity)	dis_rail_stop_km	Continuous/Categorical	km/100m
Euclidean distance to the nearest Intercity train station	dis_IC_stat_km	Continuous/Categorical	km/100m
Euclidean distance to edge of nearest highway	dis_road_poly_km	Continuous/Categorical	km/50m
Euclidean distance to edge of nearest railroad	dis_rail_poly_km	Continuous/Categorical	km/50m
Euclidean distance to the nearest highway entrance or exit	dis_hw_ex_km	Continuous/Categorical	km/50m
Location in a zone of restricted urbanization subjected to buffer-zone open-space policy	Bufferzone	Discrete (1/0)	
Location in zone of restricted urbanization subject to the Dutch National Airport Regulations Act (1996) (only for Amsterdam)	Schiphol_infl	Discrete (1/0)	
Location on the Northern side of the river IJ(only for Amsterdam, which was founded on the Southern shore)	AdamNorth	Discrete (1/0)	
Location on Southern side of the river Rhine (only for Rotterdam, which was founded on the Northern shore)	Rot_North	Discrete (0/1)	

Note: Classified as individual segments of 100 or 50 metres, from 0 to the trial distance threshold, as explained in Section 5; above the threshold the influence is considered negligible.

5. Statistical Application: Logistic Regression

This section describes the following statistical analyses performed on our extensive data set:

1. *discriminant analysis*, a technique to find out which of the selected variables is better in discriminating between any two groups of cells (in our case, built-up or not);
2. *logistic regression* using all independent variables selected as continuous variables, in order to analyse the role of each individual variable in relation to all others.

Discriminant analysis is used to find out which variables perform better in discriminating between two or more groups, whereas the logistic regression is used to identify the effects that a variable has on the probability of belonging to a certain group.

The discriminant analysis applied to the independent variables mentioned above and to the dependent variable ‘built-up or not’ led to interesting results: the Wilks’ Lambda test appears to be

significant, which means that a linear combination of the selected variables is able to discriminate between the two groups (built-up=1, built-up=0). The standardized coefficients of the discriminant function are interpreted considering that the higher the absolute value of the coefficient, the higher the contribution of the corresponding variable to the regression function. The discriminant power of an independent variable is, of course, dependent on its correlation with the other independent variables.

From Table 2 we can derive that the best linear combination that discriminates a built-up cell from a non-built-up cell, e.g. for the city of Amsterdam, is the one for which the distance to the railway polygon has the highest but negative weight, followed by the variable that indicates the presence within a zone of restricted urbanization (bufferzone), the distance from the railway stops and IC stations, the distance from the city centre, and the distance from the road polygon (the latter has a negative impact). Although the other variables (Amsterdam-North, Schiphol_infl, dis_hw_ex_km) have positive, but weak weights, we decided to keep them as explanatory variables in our logistic regression model. Since these are preliminary results, it is clear that the weight assigned to each variable is different depending on the urban context under examination.

The logistic regression technique is a statistical method commonly used in the area of social and behavioural sciences to assess the influence of several characteristics on a given phenomenon that can be represented by means of a dichotomic (or binary) variable (see, e.g., Fragkias and Seto, forthcoming). The aim of a logistic regression is the same as a linear regression, but the hypotheses at the basis of the latter method are not satisfied if the dependent variable is dichotomic (Kleinbaum and Klein, 2002; Christensen, 1997). Since the characteristic indicator (built-up or not) for which we want to examine the existence of a relationship with the accessibility indicator is a binary variable, the logistic regression is an appropriate statistical analysis to use.

Next, the application of the binomial logistic regression using “built-up or not” as the dependent variable and the nine independent variables (see Table 1) was performed. The statistical software used to perform the analyses of our case studies is SPSS, in which data processed from a GIS software (ESRI ArcInfo) are imported.

The results on which we will focus our attention in the next part of the paper are:

- the Chi-squared statistics (also known as G_M or Model L^2) indicating the level of significance: it tells us how good the model is, since, if the significance is close to 0, it means that one or more β's differ from 0, although it does not specify which ones;
- the “-2loglikelihood” goodness of fit statistic (also known as Model Deviance or DEV_M) tells us how bad the model is. The likelihood represents the probability of the observed results, given the parameters estimated. Usually, the likelihood is a small number less than 1; hence, in order to handle higher numbers it is customary to use -2 times the log of the likelihood (-2LL). A good model is one that results in a high likelihood of the observed results. This translates into a small value of -2LL. So if the model fits perfectly, the likelihood is 1, and -2LL is equal to 0.
- the Cox and Snell R Squared and the Nagelkerke R Squared are indices analogous to the R^2 statistics in the linear regression: the closer the R^2 is to 1, the better the model fits the reality.

- the classification table that describes the effective correct percentage predicted from the estimated model.
- β values and exp(β) values that are the β-values in the model with the interpretation explained in Annex A.

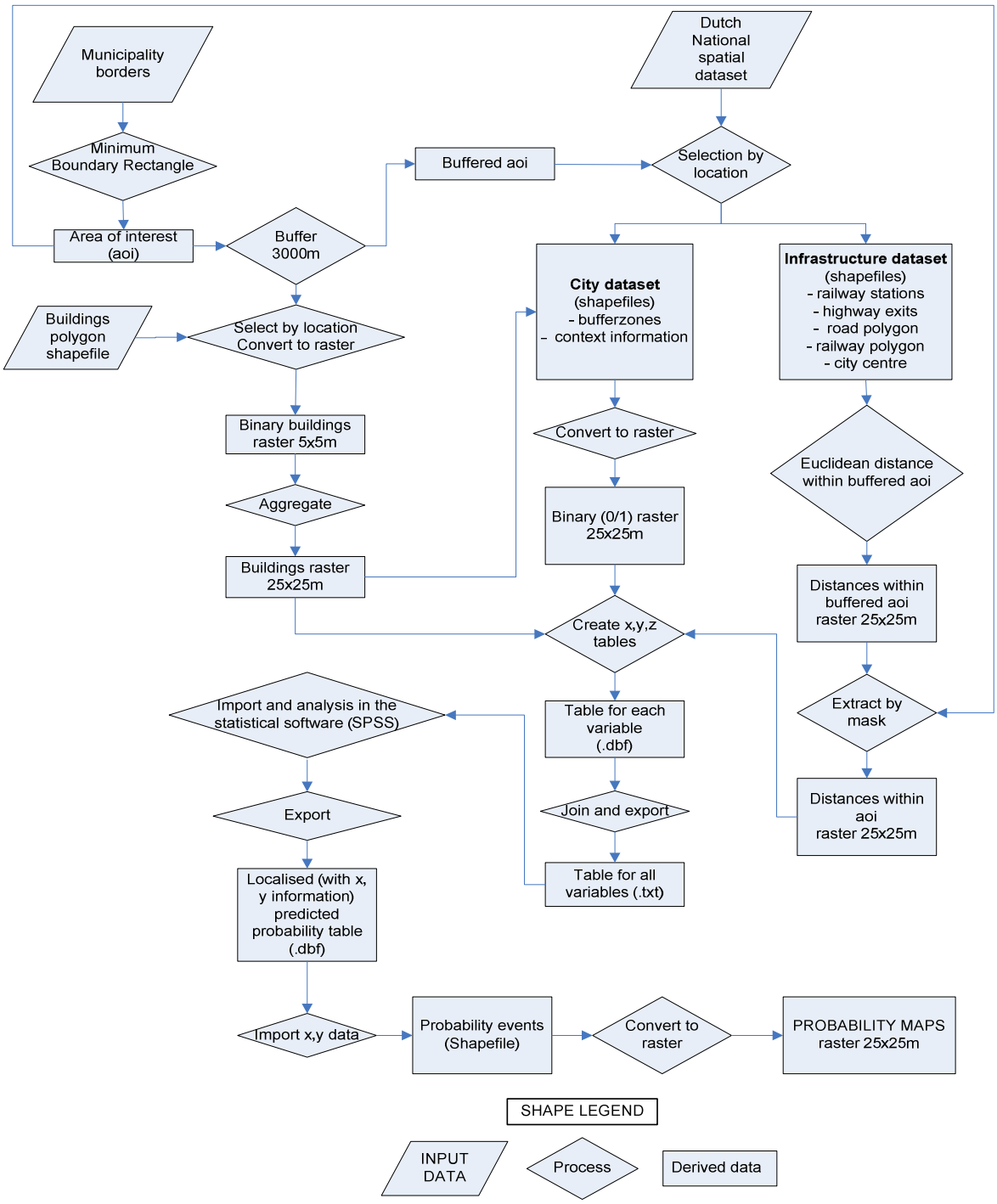


Figure 1. Data processing diagram.

Table 2. Standardized Canonical Discriminant Function Coefficients

	Coefficients			
	Amsterdam	Utrecht	Rotterdam	The Hague
Bufferzone	0.514	0.207	0.360	0.621
Schiphol_infl	0.286	--	--	--
AdamNorth	0.212	--	--	--
Rotterdam North	--	--	0.030	--
Dis_rail_stop_km	0.486	0.401	0.183	-1.653
Dis_IC_stat_km	0.472	0.702	0.813	-1.715
Dis_road_poly_km	-0.315	-0.680	-1.675	-2.652
Dis_rail_poly_km	-0.849	-0.002	-0.163	1.872
Dis_hw_ex_km	0.154	0.358	2.013	2.996
Dis_city_centre_km	0.363	-0.137	-0.278	1.703

It should be noted that, since it is not so obvious to think in terms of lnodds, in our interpretation we will consider the exp(β), in order to take into account the expected variations in the odds ratios. Simply using the definitions of the logistic regression method, and the β values from the model, the probability Prob (built-up =1) is derivable calculating Log(Odds) as:

$$Log(Odds(Y)) = \beta_0 + \sum_i^n \beta_i \cdot X_i ,$$

where X is a vector representing the n independent variables; and Y is the binary dependent variable (built-up or not), and then calculating Odds(Y) as:

$$Odds(Y) = \exp(Log(Odds(Y))$$

and eventually calculating the probability that a cell is built-up according to the regression applied:

$$P(Y) = p(\text{built - up} = 1) = \frac{Odds(Y)}{1 + Odds(Y)} \cdot 100 .$$

First, we consider the use of continuous variables in the statistical analyses performed. Then, to highlight that, beyond a certain threshold, a particular distance variable loses its role in significance within the regression function, we decided to test the results and the significance of the model by classifying five variables (distance from road polygon, distance from railway polygon, distance from highway exits, distance from regular railway stops, and distance from IC stations) as categorical variables. Thus, we had to recode the variable by assigning a constant value to the defined distance intervals, choosing a distance threshold beyond which we assume that the influence of that variable is negligible (a hypothesis to be tested by checking the significance of the regression for that value). First, the recoding was made by using a step interval of 25 m and a threshold of 400m, but in this case no meaningful differences between the subsequent β values for each step were found, while for the threshold the significance was still good. After some trial and error tests performed on data concerning the city of Amsterdam, a better combination was found by classifying the variables in 50 steps of 50 m and a threshold value of 2500m, beyond which the model significance for the three

different variables appeared to be reduced. The same “trial and error” procedure was performed for each case study, leading to the choice of different distance thresholds.

In the next section the results obtained will be compared by means of GIS probability maps.

5. Interpretation and Discussion of Results

Following the framework presented in the previous section, we will present here the most significant results, using the output for the software SPSS 13. The obtained outcomes will be explained in detail, for the four case studies (see Tables 3 and 4).

The Chi-Squared statistics were significant for each case study; Table 3 provides information about the model goodness-of-fit: the -2Loglikelihood is quite high, but the R-Squared is alright, while the overall correct percentage of the predicted values is good, viz. higher than 60 per cent for each case study. The results of the logistic regression with three categorical variables for the city of Amsterdam are reported in Annex B: distance from road polygon, distance from railway polygon, and distance from highway exits. It is worth taking a closer look at the obtained coefficients and to try to interpret them in order to derive some ideas about the distance thresholds. Considering the constant coefficient, we see that, when all the independent variables are 0, the probability of having a built-up cell is 18.315 times higher than the probability of not having a built-up cell. It is noteworthy that considering all independent variables equal to 0 implies, in this case, considering a hypothetical cell which is exactly in the city centre, 0 km away from a railway stop and from an IC station, not within the Amsterdam North zone, not within a building restriction zone, not under the influence of Schiphol airport, and more than 2500 metres away from a road polygon, a railway polygon and highway exits.

The independent variable “distance from the city centre” has a negative β of -0.107, which means that for every kilometre the lnodds decrease at a rate of (-0.107), whereas the exp(β) for that variable is 0.899: that is, holding constant all the other independent variables, for each kilometre away from the centre, the probability of having a built-up cell is 0,899 times higher than the probability of not having a built-up cell, or, in other words, the probability of having a built-up cell is 1/0.899 = 1.11 times lower than the probability of not having a built-up cell. The same interpretation must be used for the other βs: we may draw similar inferences for the independent variables ‘distance from railway stops’ and ‘distance from IC stations’, basically a type of behaviour that we could have expected.

Concerning the other six variables, that are categorical, let us take as an example the inclusion (or exclusion) in the Amsterdam-North zone. Here, the exp(β) is equal to 0.592 with a good level of significance: that is, holding constant all the other independent variables, if the cell is included in the Amsterdam-North zone, the probability of having a built-up cell is 0.592 times higher than the probability of not having a built-up cell (or, in other words, the probability of having a built-up cell is 1/0.592 = 1.69 times lower than the probability of not having a built-up cell) if compared with the ‘reference group’ made up of all the other cells not included in Amsterdam-North zone.

The same happens for the independent variable describing the influence of Schiphol: the exp(β) is equal to 0.395 with a good level of significance: that is, holding constant all the other independent

variables, if the cell is under Schiphol airport's influence, the probability of having a built-up cell is 0.395 times higher than the probability of not having a built-up cell (or, in other words, the probability of having a built-up cell is 1/0.395 = 2.53 times lower than the probability of not having a built-up cell), if compared with the 'reference group' made up of all the other cells not under Schiphol airport's influence.

Similarly, for the independent variable that represents the inclusion (or exclusion) within a building restriction zone, the exp(β) is equal to 0.097 with a fair level of significance: that is, holding constant all the other independent variables, if the cell is included in a restricted zone, the probability of having a built-up cell is 0.097 times higher than the probability of not having a built-up cell (or, in other words, the probability of having a built-up cell is 1/0.097 = 10.3 times lower than the probability of not having a built-up cell) if compared with the 'reference group' made up of all the other cells not included in the restricted zone.

Concerning the other categorical variables, which are made up, for example, of 51 groups (for the regression analyses performed assuming a threshold of 2500 m), let us take as a first example the dummy variable related to the distance from a road polygon: the exp(β) value for the first group (that is the group of all cells with a distance from road polygon between 0 and 50 m) is equal to 0.057 with a good level of significance. This means that, holding constant all the other independent variables, if the cell is included in this group, the probability of having a built-up cell is 0.057 times higher than the probability of not having a built-up cell (or, in other words, the probability of having a built-up cell is 1/0.067 = 17.54 times lower than the probability of not having a built-up cell), if compared to the 'reference group' made up of all the other cells that have a distance to the road polygon of over 2500 metres.

The same holds for all other variable groups included in this dummy variable: we see that, for each group of 50 metres, holding constant all the other independent variables, the variation of probability of having a built-up cell as opposed to not having one is always higher than the previous group; if compared with the 'reference group', then the coefficients are slightly unstable between 350 m and 400 m, but they rise again till 2000 m, in which case the significance is no longer good: beyond this threshold the variable is no longer able to explain the dependent variable. The same behaviour is found by looking at the exp(β) coefficients of the independent variable 'distance from railway polygon', with the difference that there is a good level of significance up till 2450 metres.

The independent variable 'distance from highway exits' has the same interpretation, but we find some differences in the figures. First of all, the exp(β)s are all higher than 1 or very close to 1, while for figures approaching 1, the result does not show a satisfactory significance. For the latter groups (with exp(β) approaching 1) therefore, there are no significant differences with respect to the reference group (the group with a distance from the highway exits exceeding 2500 metres), whereas for those groups of cells at a distance 0 to 2000 metres from highway exits, the significant exp(β) around 2 means that the probability of having a built-up cell is 2 times higher than the probability of not having a built-up cell. This sounds plausible, since it is unlikely that there will be a built-up cell immediately close to a highway exit, whereas it is conceivable that activity functions may be located almost near a highway exit, and certainly not very far away.

Thus we expect to have exp(β) lower than 1e for the variables bufferzone, Amsterdam North, Schiphol influence, Rotterdam North; lower than 1 but increasing with the distance when categorical for the variables railway polygon and road polygon distances; higher than 1 initially increasing then decreasing for the variables highway exits and railway and IC stations distances.

An overview will now be provided of the various model coefficients of the logistic regression performed with all continuous variables for the four Dutch case studies and the indication of the accessibility threshold found, along with the input map of each case study and the final probability map.

While Table 3 is able to tell us about the goodness of fit of the model, the probability maps represented in Figures 2-5 describe the probability that each cell is built-up according to the model prediction, taking into account therefore each of the independent variables. Table 4 describes the coefficients derived from the logistic regression using only the continuous variables, whose values are the ones we expected.

Table 3. Summary of goodness-of-fit statistics of the logistic regression performed with all continuous variables

Statistics	**Case study**			
	Amsterdam	**Utrecht**	**Rotterdam**	**The Hague**
-2 Log likelihood	379255	138923	653391	230893
Cox & Snell R Squared	0.147	0.179	0.107	0.123
Nagelkerke R Squared	0.205	0.244	0.158	0.165
Overall correct model percentage	72.1	71.8	74.5	62.7

Table 4. Summary of the logistic regression coefficients performed with all continuous variables

	Amsterdam		**Utrecht**		**Rotterdam**		**The Hague**	
VARIABLES	β	Exp (β)	β	Exp(β)	β	Exp(β)	β	Exp(β)
Bufferzone	-1.672	0.188	-0.801	0.449	-2.354	0.095	-1.777	0.169
dis_rail_stop_km	-0.259	0.772	-0.315	0.730	-0.121	0.886	0.714	2.042
dis_IC_stat_km	-0.272	0.762	-0.747	0.474	-0.122	0.885	0.628	1.874
dis_road_poly_km	0.438	1.550	0.504	1.656	0.328	1.388	0.922	2.514
dis_rail_poly_km	0.560	1.750	-0.060	0.942	0.165	1.179	-0.782	0.457
dis_hw_ex_km	-0.318	0.727	-0.087	0.917	-0.498	0.608	-0.996	0.369
dis_city_centre_km	-0.106	0.899	0.289	1.335	-0.057	1.058	-0.611	0.543
Schiphol_infl	-0.775	0.461	--	--	--	--	--	--
AdamNorth	-0.423	0.655	--	--	--	--	--	--
RotterdamNorth	--	--	--	--	-0.197	0.821	--	--
Constant	0.844	2.325	1.546	4.692	0.282	1.325	0.974	2.648

From Table 5 we see that the larger the study area, the bigger the threshold value for each type of infrastructure. Comparing the input maps, we can also see that the number, the location and the distribution of the infrastructures are parameters which can explain the differences between the

thresholds: remarkable differences can be observed between the case study of Utrecht (for which anyway the distance from the railway polygon is found to be a non-significant explanatory variable), which has a small but uniformly distributed centre, and the city of Rotterdam, whose big threshold values are due to the large size of the area concerned and to the particular distribution of the infrastructures.

The probability maps shown in Figures 2-5 are used to facilitate the interpretation of the results originating from the model coefficients: in each map, comparing them with the input variable maps, and in particular with the buildings, we can see that the actual localization of the built-up areas can be considered to be predicted well by the model.

Table 5. Estimated distance thresholds compared with the extent of the built-up areas for each case study

	Amsterdam	**Utrecht**	**Rotterdam**	**The Hague**
built-up areas extent	70.216 km^2	28.929 km^2	103.105 km^2	54.913 km^2
total area of interest*	224 km^2	80.99 km^2	408.233 km^2	116.970 km^2
road polygon threshold	2000 m	1000 m	4000 m	1200 m
railway polygon threshold	2500 m	not significant	> 4500 m	3600 m
highway exits threshold	2000 m	500 m	4000 m	300

*For the cities of Rotterdam and The Hague the total area of interest does not consider the no data cells.

6. Retrospect and Prospect

The first aim of this study was to quantitatively explain the presence of actual built-up areas and their spatial position in the city by means of accessibility indicators and a limited set of other spatial variables. The probability maps and the model coefficients reported in the previous section show that by means of our logistic model, we are able to generate localized probability predictions of the presence of built-up areas, which explains the actual pattern with a rather good level of significance. We found that, by means of the discriminant analysis, that the order of importance of the selected variables in explaining built-up areas as follows: first the distance from the city centre, then the presence of planning or environmental constraints, followed by the proximity of railway stations and highway exits.

The second aim was to find a quantitative and scientific methodology to define the threshold of influence of different types of infrastructure on the loss of significance of the relevant categorical variables with increasing distances. We found that the methodology proposed appears to be a useful and reliable tool to assess these thresholds. The obtained results are different, depending on the type of infrastructure and the urban space, and especially on the size of the area of interest. This is the size of the minimum bounding rectangle of the municipality borders, very close to the size of the town itself. The finding confirms our expectations and can be related, for example, to the studies about the influence of transportation infrastructures on land prices: for example, the impact of the railway is found to have a declining influence with the distance, up to 10,000 km (Debrezion, 2006). These outcomes also offer useful input to other studies that deal with the modelling of future land-

use patterns and rely on accessibility indicators to define the suitability for specific land-use types (Koomen et al., 2007).

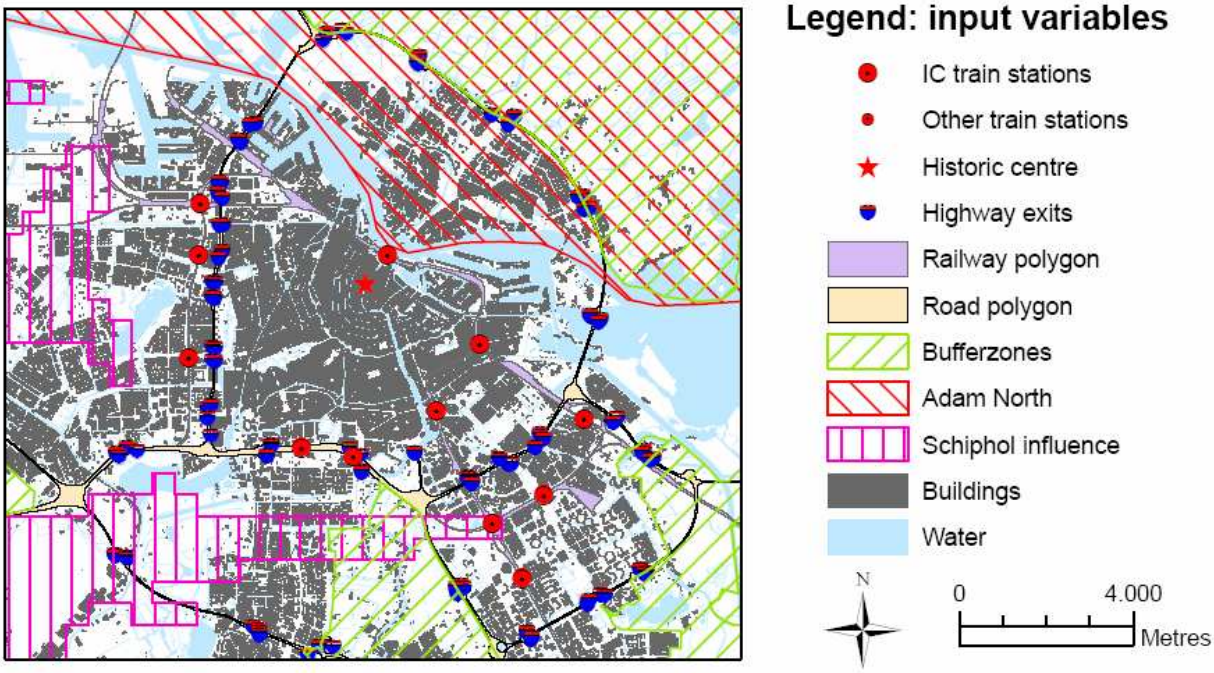
Another issue to be explicitly considered is that our experiments do not provide a comprehensive study, since they focus only on urban areas, while neglecting rural areas. Moreover, many other variables could be considered as explanatory for the localization of built-up areas. Furthermore, we used as an accessibility indicator only Euclidean distances and a spatial scale of 25m by 25 m; further experiments using network distances, or different measures of accessibility are needed, in order to take into account, for example, even the influence of travel behaviour (Geurs et al., 2006).

There is also the fundamental issue of the data interoperability to be highlighted: without interoperability between different software, the duration and the processing weight of the analyses would have been very high. All the results would have not been observed without the flexibility and the interoperability between different software such as GIS and statistical ones: these modern software tools can be of great help to knowledge based urban and transportation planning.

It should also be stressed that, to perform meaningful statistical analyses on the transportation and land-use systems it is important to collect data and systematically monitor them. The statistical analyses under study are meaningful only if there is a good sample of data on which perform them. This consideration is clear if we look at the case study of The Hague, for example, for which the influence of highway exits is not significant since there are no data on highway exits in the city (see the grey areas in Figure 5 on the outskirts of The Hague). This contribution is part of the larger body of research that tries to explain the relationships between the land use and transportation system: it only gives a methodology able to deal with problems such as the choice of accessibility threshold, but more study is needed to highlight the dynamic relationships between these two systems.

Starting from the simple indicators selected, a further research subject could, for example, be the investigation of the same relationships but choosing different time periods, for the built-up are dependent variable and the infrastructures and constraints variables, in order to consider the well-known chicken-egg problem: Does infrastructure explain urbanization, or does it merely follow?

Complex land use/transportation models have been built in recent decades to answer this question in a detailed manner (Geurs et al., 2006); these quantitative analyses could be useful in order to screen at a higher and macro-level which are the most meaningful variables to take into account and the direction of relationships. Eventually, it would be interesting to use the same methodology to investigate this relationship for other major European cities, in order to place these findings in an international context.



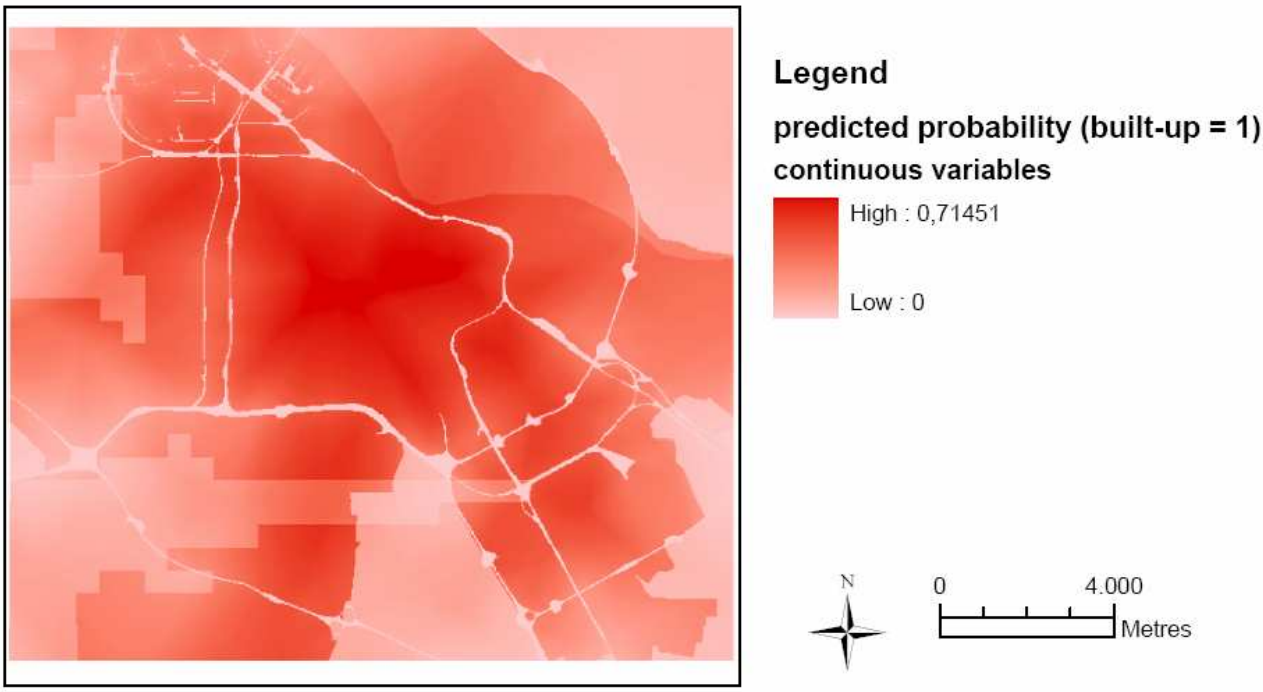


Figure 2. Study area and probability map - Amsterdam

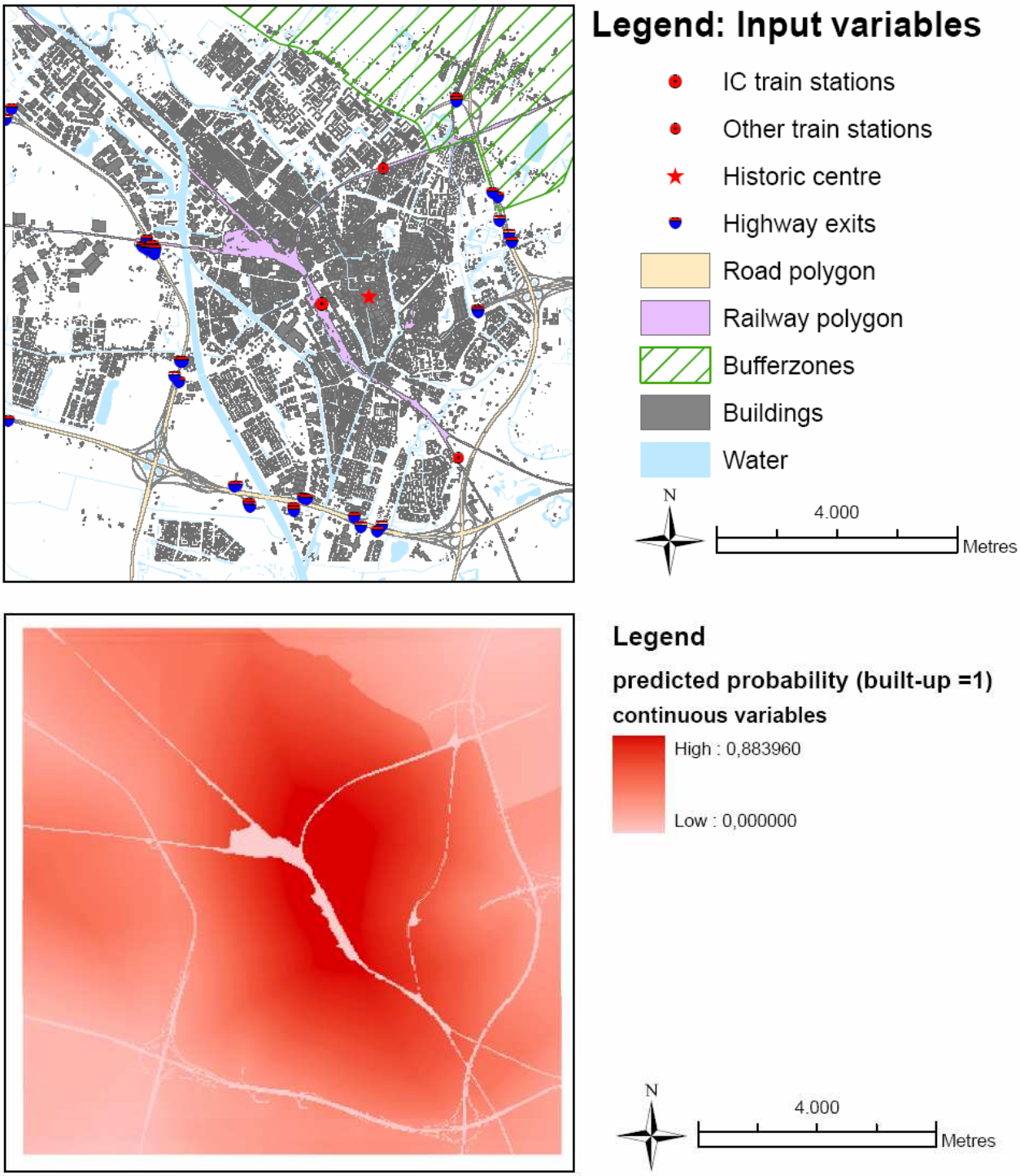


Figure 3. Study area and probability map – Utrecht

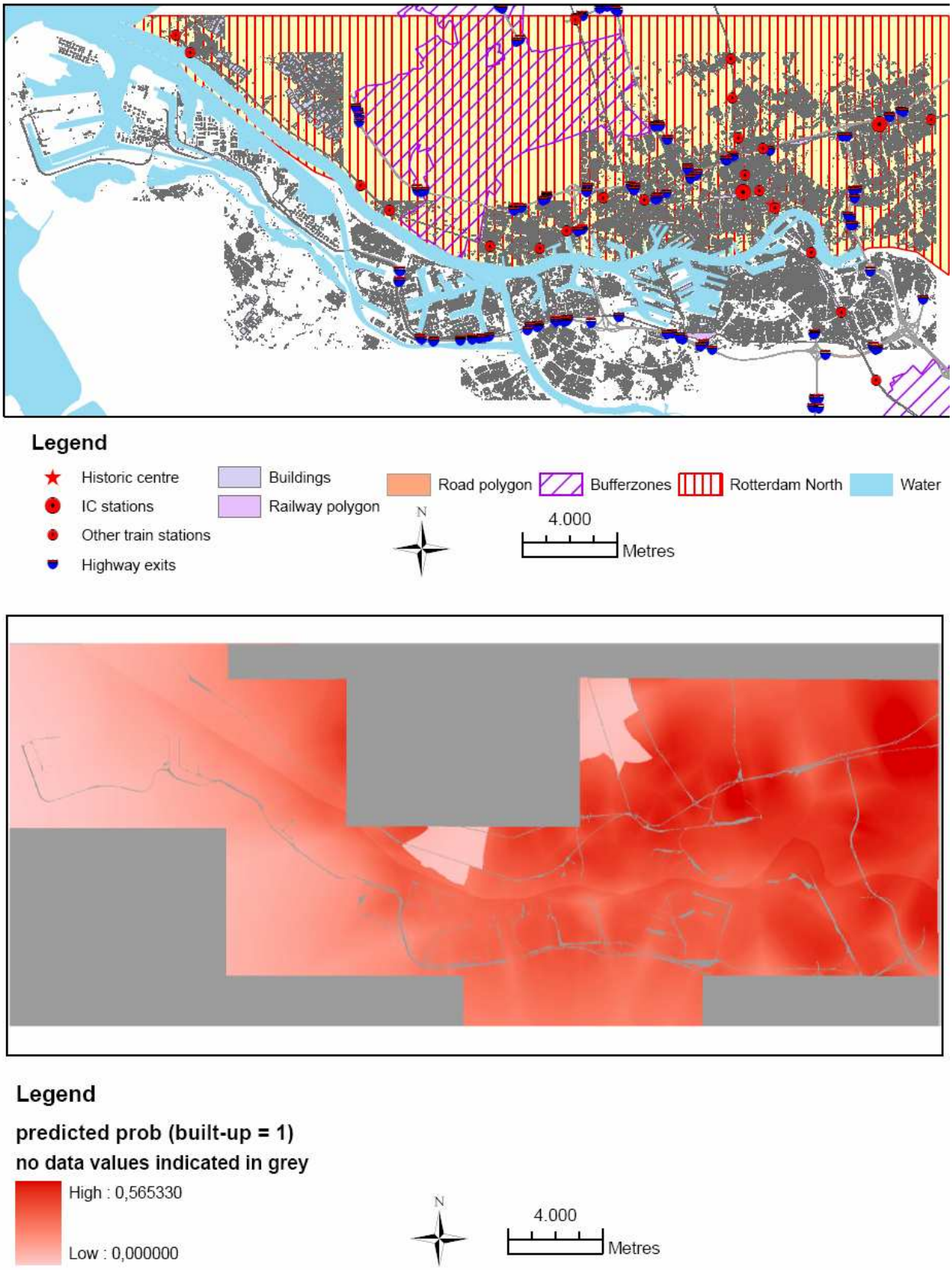


Figure 4. Study area and probability map - Rotterdam

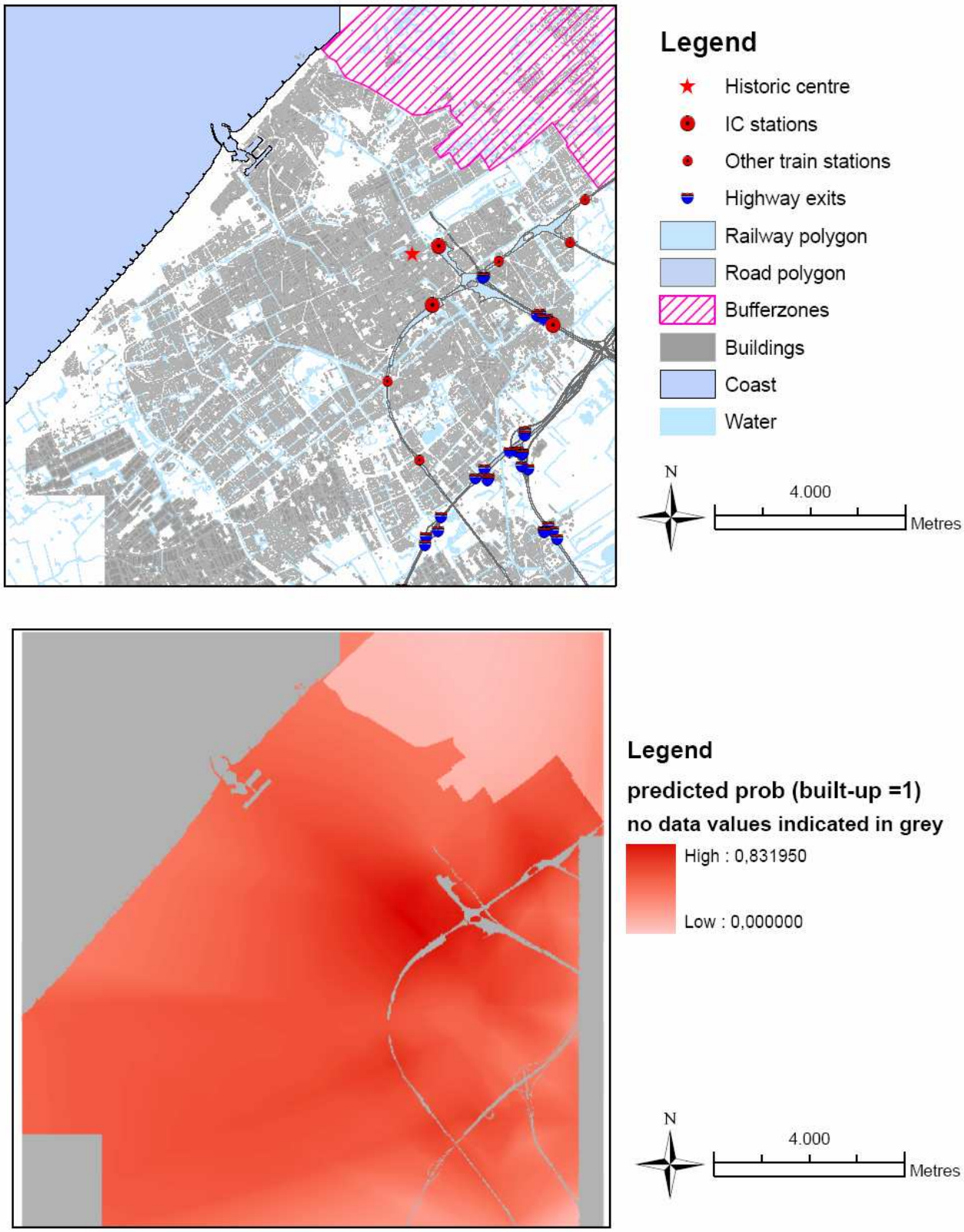


Figure 5. Study area and probability map – Den Haag

References

Adler, E.S., Forrest, N. (1984), *Linear Probability, Logit, and Probit Models*, Quantitative Applications in the Social Sciences Series, SAGE Publications, Delhi.

Blunden, W. R. (1971), *The Land-Use /Transport System: Analysis and Synthesis*, Pergamon Press, New York.

Christensen, R. (1997), *Loglinear Models and Logistic Regression* (2nd edition), Springer-Verlag, Berlin.

de la Barra, T. (1989), *Integrated Land-Use and Transport: Decision Chains and Hierarchies*, Cambridge University Press, Cambridge.

Debrezion Andom, G. (2006), Railway Impacts on Real Estate Prices, PhD Thesis, Tinbergen Institute Research Paper no. 389, Thela Thesis, Amsterdam.

Dekkers, J., Koomen, E. (2007), Land Use Simulation for Water Management, *Modeling Land Use Change* (E. Koomen, J. Stillwell an H.J. Scholten, eds.), Springer Verlag, Berlin, pp. 355-373.

Ettema, D., Timmermans, H. (2007), Space-time Accessibility under Conditions of Uncertain Travel Times: Theory and Numerical Simulations, *Geographic Analysis*, vol. 39, no. 2, pp. 217-240.

Fragkias, M., Seto, K. C. (2007), Modeling Urban Growth in Data-sparse Environments: a New Approach, forthcoming in *Environment and Planning B*.

Geurs, K. T., Ritsema van Eck, J. R. (2001), *Accessibility Measures: Review and Applications*, RIVM report 408505 006, National Institute of Public Health and the Environment, Bilthoven.

Geurs, K. T., van Wee, B. (2004), Accessibility Evaluation of Land-use and Transport Strategies: Review and Research Directions, *Journal of Transport Geography*, vol. 12, pp.127-140.

Geurs, K. T., van Wee, B., Rietveld, P. (2006), Accessibility Appraisal of Integrated Land-use –Transport Strategies: Methodology and Case Study for the Netherlands Randstad Area, *Environment and Planning B*, vol. 33, no. 5, pp.639-660.

Kleinbaum, D.G., Klein, M. (2002), *Logistic Regression – a Self Learning Text* (2nd edition), Springer-Verlag, Berlin.

Koenig, J. G. (1980), Indicators of Urban Accessibility: Theory and Application, *Transportation*, no. 9, pp. 145-172.

Koomen, E., Stillwell, J., Scholten H.J. (2001), *Modeling Land Use Change*, Springer Verlag, Berlin.

Martellato, D., Nijkamp, P. (1999), The Concept of Accessibility Revisited, *Accessibility, Trade and Locational Behaviour*, (A. Reggiani, ed.), Ashgate, Aldershot, pp. 17-40.

Martellato, D., Nijkamp, P., Reggiani, A., (eds.) (1995), Measurement and Measures of Network Accessibility: Economic Perspectives, *Transport Networks in Europe: Concepts, Analysis and Policie*s (K. Button, P. Nijkamp and H. Priemus, eds.), pp. 161-180.

Martin, J. C., Reggiani, A. (2007), Recent Methodological Developments to Measure Spatial Interactions: Synthetic Accessibility Indices Applied to High Speed Train Investments, forthcoming in *Transportation Review*.

Reggiani, A. (ed.) (1998), *Accessibility, Trade and Locational Behaviour*, Ashgate, Aldershot.

Rietveld, P., Bruinsma, F. R. (1998), *Is Transport Infrastructure Effective?*, Springer-Verlag, Berlin.

Schoemakers, A., van der Hoorn, T. (2004), LUTI Modeling in the Netherlands: Experiences with TIGRIS and a Framework for a New LUTI Model, *European Journal of Transport and Infrastructure Research*, no. 3, pp. 315-332.

Timmermans, H.J.P. (2003), The Saga of Integrated Land Use - Transport Modeling: How Many More Dreams Before We Wake Up? *Proceedings of the International Association of Travel Behaviour Research*, Lucerne, August 10-15.

Tomlin, D. (1990), *Geographic Information Systems and Cartographic Modelling*, Prentice Hall, New Jersey.

Zondag, P., Pieters, M. (2005), Influence of Accessibility on Residential Location Choice, *Journal of the Transportation Research Board*, no. 1902, TRB, Washington DC, pp. 63-70.

Annex A: Logistic regression interpretation

As mentioned above, the choice of the logistic regression is explained by the particular kind of data we are dealing with, that is, they are not suitable for a linear regression. Since we want to investigate the relationships between a dichotomic dependent variable, i.e. the presence or absence of a built-up area, and a group of several quantitative independent variables, the linear regression is not suitable because the following assumptions are problematic:

1. a linear relationship between dependent variable and independent variables;
2. homoschedasticity for the errors (that is, the error variance is constant)
3. normal distribution for the regression errors, in testing the model significance.

If these assumptions are violated, the following mistakes are made:

1. if the relationship is not linear, our model is misspecified;
2. if, for different values of the independent variables, the error variance is different, it is not possible to consider only one value of R^2 in order to interpret the model; it would lead to several mistakes in the description of the explained variance;
3. if the errors are not normally distributed, the significance test is incorrect.

Such assumptions are simple to meet if the dependent variable is continuous. But with a dichotomic (or categorical) dependent variable there are some caveats:

1. the linearity condition is not satisfied, because the relationship between the variables is concentrated around the only two values of the dependent variable;
2. the homoschedasticity is not satisfied, because the variable is dependent on the predicted value
3. the normal distribution of the errors cannot be achieved; for example, if we consider that the error is the difference between observed and predicted values - when the possible observed values are 0 or 1, while the predicted values are probabilities - the error distribution will allways be bimodal, with two humps, very different from the normal one.

As a solution, the logistic regression (Adler et al., 1984; Christensen, 1997; Kleinbaum and Klein, 2002) aims to work in terms of a linear relationship, by simply carrying out a change in variables, by extending the domain between $-\infty$ and $+\infty$, and using *odds ratios*.

The first rationale on which the logistic regression is based is to predict not the probability of occurrence of the dependent variable (e.g. P[built-up=1]), ranging between 0 and 1, but the odds ratio P/(1-P), that is the ratio between the probability of the event's occurrence and the probability of the event's non-occurrence. The odds ratio indicates how much an event is more probable with respect to its complementary event. The odds ratio ranges between 0 and $+\infty$, and has the following functional relationships with probabilities (see Table A.1).

Since the odds have a lower bound of 0, and a linear function can have a range in the total real number space, it is possible to eliminate it by means of the logarithm of the odds ratio, also known as the logit: ln(P/(1-P))=logit(Y), which ranges between $-\infty$ and $+\infty$. The specific aim of a logistic

regression analysis is to estimate the logit function. Table A.2 can help in understanding the effect on the probability estimation.

Table A1. Relationships between probabilities and odds ratios

Probability	Odds ratio	Explanation
0.5	1	Equiprobable events
>0.5	>1	The event with p>0.5 is more probable than its complement
<0.5	<1	The event with p<0.5 is less probable than its complement

Table A2. Relationships between probabilities, odds ratios and logits

Probability	Odds ratio	ln(odds)	Explanation
0<p<0.5	0<odds<1	>0	The event with p<0.5 is less probable than its complement
0.5	1	0	Equiprobable events
>0.5	>1	<0	The event with p>0.5 is more probable than its complement

The graphical relationships between the presented functions - according to the model assumptions – imply that the logit has a sigmoid relationship with the probability, whereas the independent variable has a linear relationship with the logit.

Hence, when a logistic regression is performed, we are assuming that the following function is able to fit our data:

$$\ln(Odds(Y)) = \beta_0 + \sum_i^n \beta_i \cdot X_i$$

where Y is the dichotomic dependent variable; and X_i are the *n* independent variables; the aim of the regression is now to estimate the β_is that can reproduce the observed data. In terms of odds ratios, the same relationship can be written using a simple exponential transformation:

$$Odds(Y) = e^{\beta_0 + \sum_i^n \beta_i \cdot X_i} = e^{\beta_0} \cdot e^{\sum_i^n \beta_i \cdot X_i} = e^{\beta_0} \cdot \prod_{i=1}^n \left(e^{\beta_i}\right)^{X_i}$$

The results of the logistic regression are therefore estimates of the βs or the exp(β)s, depending on the regression function on which we are focusing. The βs may be interpreted as the expected change in the logit when the independent variable X increases by one unit. The β_0 value, corresponding to the constant coefficient, is the expected change in the logit, when all independent variables are set to null. The exp(β) is the rate of increase in the odds ratio for each increasing unit of the independent variable.

If a multiple logistic regression is applied using more than one independent variable, the exp(β) values (as well as β values) cannot be regarded as the odds variation with respect to a unit increase of the corresponding X, because the correlation with the rest of the independent variables must be taken into account. Besides, in this case, the coefficient exp(β) (or β) related to the constant term represents the expected logits (or odds) when all independent variables are set to null. The

coefficient exp(β) (or β) related to any particular X is the expected variation of logit (or odd) for a unit increase of X, when all other independent variables are held constant.

In other words, the single β_i in the multiple logistic regression indicates:

1. the effect of X_i, net of the effects due to X_j ($j=1\ldots n$, $i \neq j$);
2. the expected variation (in logit) for a unit increase of X_i, removing the variability due to X_j ($j=1\ldots n$, $i \neq j$) and every relationship between X_i and X_j.

There are some cases (for example, when there is a reference group set as a target group) in which we may want to define a 'categorical' independent variable. This means that the specific independent variable is subdivided into more than two groups: the categorical variable is represented as a series of dichotomic variables that represent the differences between groups. For example, an independent variable with k groups, is represented by means of k-1 dummy variables in a k-dimensional matrix, whose rows have 1 in a specific position for each independent variable group, except one of them, called the 'reference group' that is represented only by a '0' row. Using categorical independent variables, the logistic regression thus estimates the effects on the dependent variable of all the independent variable with respect to the reference group.

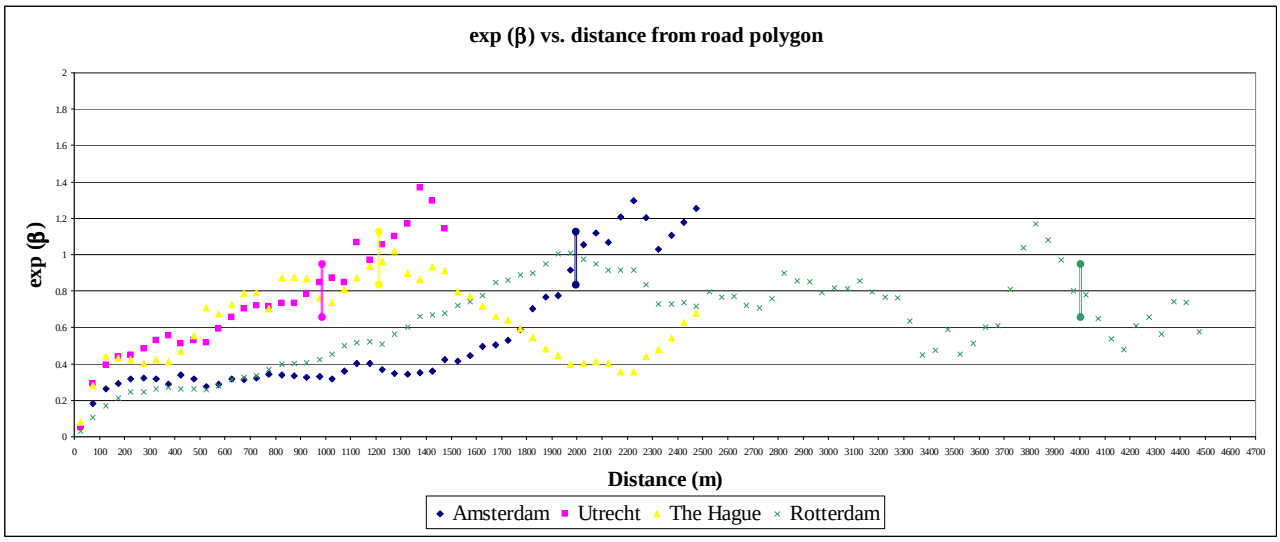
Therefore, we may conclude regarding the results of the logistic regression with categorical independent variables:

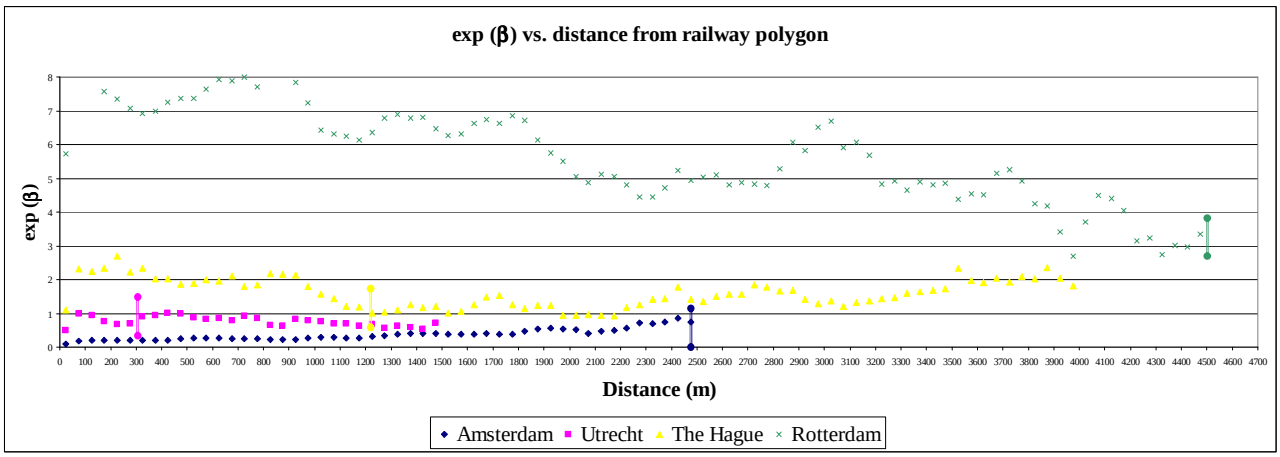
1. the constant term represents the expected variation of the odds ratio for the reference group, holding all other groups as 0;
2. the other coefficients represent the variations of odds ratios passing from the reference group to the group corresponding to that dummy variable;
3. there is no coefficient present for the overall model, because the effects vary from group to group: every exp(β) is related to the reference group, that is, it is an $\exp(\beta_{rif\text{-}Xi})$.

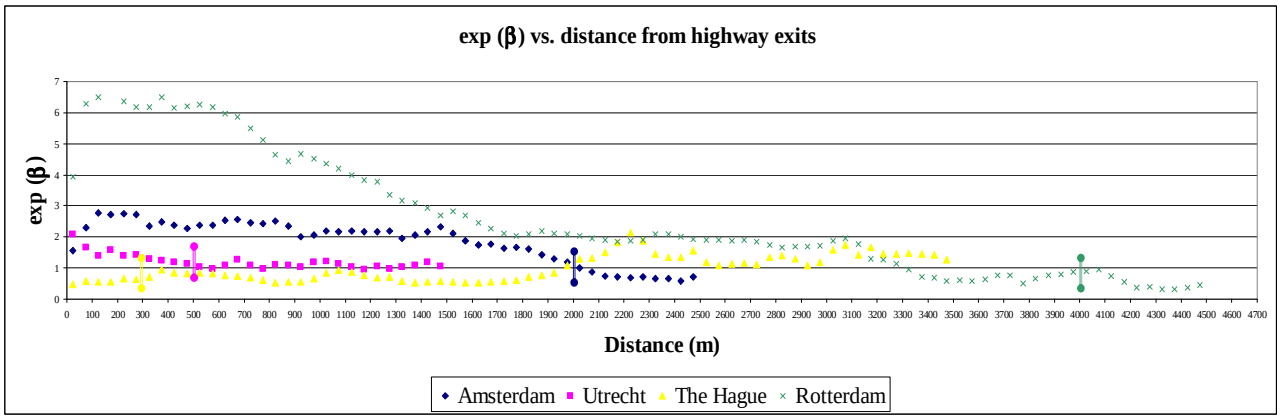
Annex B: Statistical results

The statistical results are shown in the following diagrams by means of the exp(β) coefficient and the three categorical variables (distance from road polygon, distance from railway polygon, distance from highway exits): as expected, the first two are for lower distances less than 1, and then globally increasing, meaning that the probability al having a built-up cell as opposed to not having it is higher far from a road and railway polygon. For the third variable the trend is different, the coefficients are generally more than 1, first increasing, then decreasing with the distance. The only deviations from this trend happen to the variable distance from the railway polygon for the city of Rotterdam, which may be because the railway polygon is included deep in the city, and the variable 'distance from highway exits' for the city of The Hague that is not decreasing but stable with the distance: far from highway exits there is no difference for the probability of having a built-up cell as opposed to not having it. This is understandable if we look at the city map: most of the city is developed far from highway exits, that are not central at all.

In the diagrams below the vertical lines with the same colour as the series correspond to the threshold beyond which the influence of that variable on the accessibility could be considered negligible, as previously specified in Table 5.






exp (β) vs. distance from highway exits
exp (β)
Distance (m)
Amsterdam
Utrecht
The Hague
Rotterdam