Building SIENA, a simulation for environmental health analysis

2010

By

Daniela Fecht

Department of Epidemiology and Biostatistics

Imperial College London

Building SIENA, a simulation for environmental health analysis

1. Introduction

The **SI**mulation for **EN**vironmental health **A**nalysis (SIENA) is a representation of an urban area to perform simulations in an environmental epidemiological context. SIENA provides a controlled, simplified urban environment to develop and test spatial epidemiological concepts and models, to simulate processes and interactions relating to environmental exposure and to explore theoretical and methodological problems in the spatial analysis of environmental health. The simulated urban area should typify a medium-sized city in Great Britain to allow conclusions to be drawn for real-world settings. The data structure of SIENA should reflect the real-world urban structure accordingly. Special focus will be given to the interactions between the urban components. A statistical analysis of urban components in twelve sample cities in Great Britain has shown that the complexity of the urban interactions is immense. The challenge in building an urban simulation is, therefore, to preserve as much complexity as possible. Interdependencies between the urban components are, consequently, a key focus in the modelling process. The preservation of these interactions also has to be assured if SIENA is used to model environmental epidemiological scenarios. For this purpose, the urban simulation should allow enough flexibility to simulate different scenario specific data, considering the spatial urban structure of SIENA.

These simulation requirements are reflected in the data structure of SIENA. The simulation is made up of three different data types, outlined in Figure 1. The first data type is core data, the urban structural elements topography, transportation network, land cover and population distribution. Core data is based on real-world data of the sample cities and is modelled for the urban simulation following the design rules derived from the statistical analysis of the urban components. This data is the foundation of SIENA. The second data type is contextual data. Contextual data is also based on real-world data and is modelled onto the spatial structure of the core data. It is scenario specific data and added to the simulation to further characterise the urban simulation and enhance its functionalities. Contextual data can be added to the simulation to support the case studies carried out. The third data type is derived data. Utilising both core and contextual data the simulation provides the data structure to derive new data using both established models and newly developed models. Methods and approaches used to model the different data types for SIENA are outlined below.

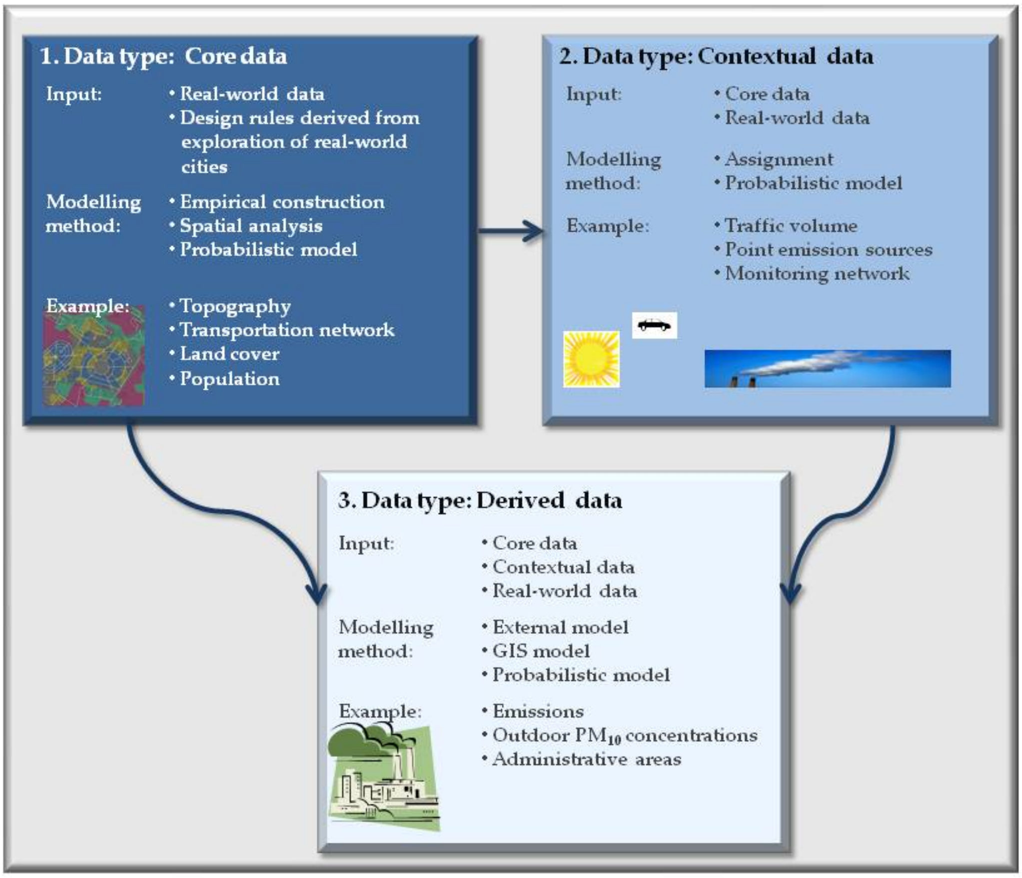


Figure 1. The SIENA data model

2. *The model premise*

Prior to modelling the three different data types, fundamental principles of the simulation environment need to be defined. The type of city to be simulated, coastal or inland city, as well as the spatial resolution and the dimensions of the urban simulation have to be established.

Inland versus coastal city

The statistical analysis of the twelve sample cities has shown that coastal and inland cities differ in their topographical characteristics. The location of the city centre also varies between the two city types. A first point to consider before starting the modelling process is, therefore, whether SIENA should be a representation of a coastal or of an inland city. It was decided to simulate an inland city. Inland cities are the dominant city type in Great Britain and are considered demographically, climatically and economically more stable than coastal cities (*Abraham & Hendershott, 1996; Balk et al., 2008; Kuttler, 2008).* The design and structural rules are, therefore, informed by the six inland sample cities: Coventry, Derby, Leicester, Nottingham, Reading and Sheffield.

Spatial resolution

The spatial resolution of the urban simulation is another aspect that needs to be defined prior to the modelling process. The choice of the resolution has to be a balance between the highest possible spatial resolution to allow for the detection of fine spatial patterns in the urban environment and the computational feasibility. In order to establish the optimal spatial resolution, tests are carried out using various resolutions ranging from 10 metres to 100 metres. For this purpose the study area of Leicester is used, the city closest to the inland sample city average in terms of city size and number of land cover patches. Five different sets of gridded points are generated for Leicester with resolutions of 10 x 10 m, 25 x 25 m, 50 x 50 m, 75 x 75 m and 100 x 100 m. A point-in-polygon analysis overlaying the land cover polygons with the gridded points identifies for each set of gridded points the land cover patches that have one or more gridded points within their boundaries. The number of identified land cover patches as well as the associated patch area is calculated as a measure of performance. As expected, a 10 x 10 metres resolution gives the best performance but proves computationally too demanding for general application. Use of a 25 x 25 metres resolution, however, greatly reduces the computational demands, while causing only a marginal reduction in performance; 80% of the land cover patches which represents 98% of the land cover patch area (see Figure 2) are identified by the point-in-polygon analysis with the gridded points.

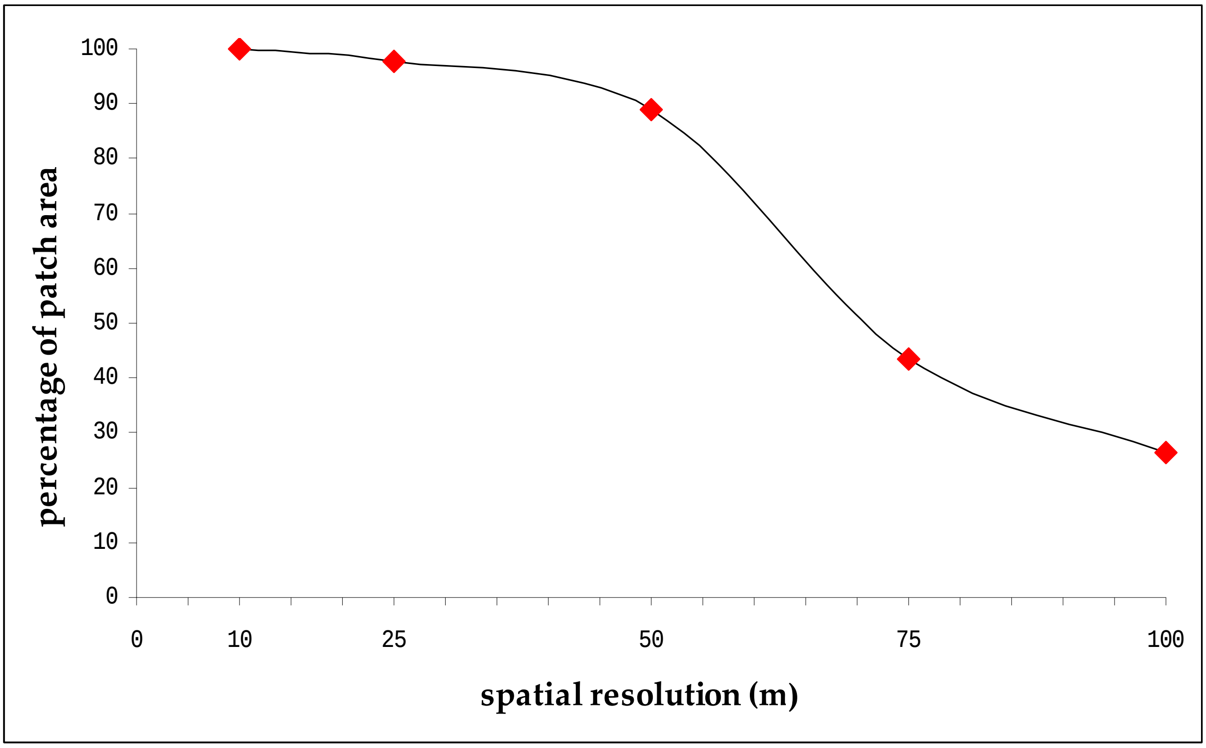


Figure 2. Percentage of land cover patch area being identified by different spatial resolutions

Twenty-five metres is also the spatial resolution of the land cover data. An increase in resolution beyond 25 metres can, therefore, not be justified because such false precision would distort interpretation of any simulated results. Spatial data for SIENA is, therefore, modelled with a 25 x 25 metres resolution.

Dimensions of SIENA

The average extent of the inland sample cities defines the dimension of SIENA. The inland sample cities encompass on average an area of 158 km^2 with average side lengths of 13.5 km and 11.7 km. The boundaries of the urban simulation environment are based on these specifications and consequently result in an area of 153 km^2 with side lengths of 13.5 km times 11.3 km. The slight discrepancy between the mean extent of the inland sample cities and the urban simulation is due to the 25 x 25 metres resolution of the urban simulation.

3. The construction of SIENA

3.1 Core data

The urban components topography, transportation network, land cover and population, make up the core data of SIENA. They build the foundation in structure and content on which later the contextual and derived data can be based. These four components provide important information for environmental health analysis in an urban area. The environmental factors topography, transportation network and land cover have a strong influence on the spatial distribution of urban pollutants such as traffic related air pollution or pesticides (*Alavanja et al., 2007; Rosenlund et al., 2008*). They also determine both directly and indirectly the distribution of the urban population via the accessibility of the terrain and the road network, they influence people's activity patterns and as a consequence affect their exposure levels (*Zuurbier et al., 2010)*.

The findings of the urban exploration of these four components in the statistical analysis are consequently translated into design rules to inform the construction of the core data structure for SIENA. Depending on the conditions formulated in the design rules and the nature of the data, different approaches are used to model the core data. Applied methods range from empirical construction of structural data, to probabilistic modelling, to the use of spatial analysis tools. Figure 3 illustrates the different approaches used to model the core data and maps the influence of one type of core data on the spatial distribution of the next in the modelling hierarchy as outlined in the design rules. This hierarchy is also maintained in the following description of the modelling approaches.

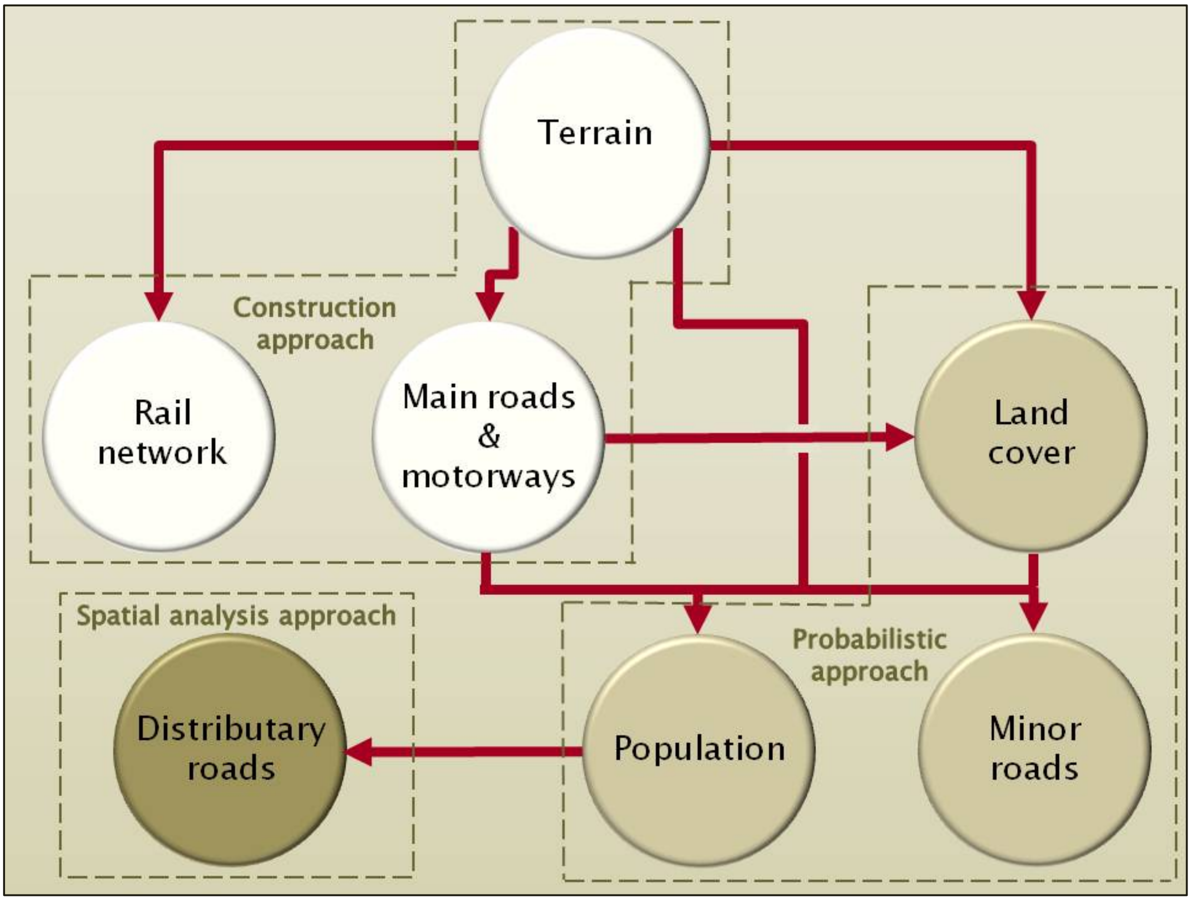


Figure 3. Modelling approaches used to model the core data for SIENA

Formulation of design and structural rules for the construction of SIENA

Based on the findings of the statistical exploration of the four urban components, topography, transportation network, land cover and population, design and structural rules are established for the modelling of these data sets. The derived rules are based on the characteristics of the inland sample cities to build a 'typical' representation of a medium-sized urban area in Great Britain.

The topography has been identified in the analysis of the urban components as the determining structural feature that strongly influences the spatial distribution of the other urban components. Interdependencies with the topography, therefore, have to be reflected in the construction rules of all other urban components.

The rules for the construction of the topography are based on the assumption that topographical features can be averaged across the inland sample cities and that the use of these average values to model the SIENA topography will result in topographical characteristics similar to the inland cities. Straightforward averages of the characteristics are therefore used as rules, focusing on the minimum and maximum altitude values as well as the altitude range. Rules for slope angle are not formulated because the slope is a direct consequence of the modelled topography.

The rules for the construction of the main transportation network follow the same assumption. Again, average values concentrating on the structural characteristics of the inland sample cities' main transportation routes are used to inform the construction of the SIENA main transport routes. The associations with topography identified in the statistical analysis have to be reflected in SIENA and the transport network will therefore be constructed after the topography. The rules for the construction of the transport network are summarised in Table 1.

Table 1. Design and structural rules

	Measure	Specifications
Topography	Minimum altitude	Within inland sample city range and close to mean
	Maximum altitude	Within inland sample city range and close to mean
	Altitude range	Within inland sample city range and close to mean
Main road network	Length of main roads	Within inland sample city range and close to mean
	Length of motorway	Within inland sample city range and close to mean
	Number of feeder roads	Mean
	Number of ring roads	Mean
	Number of distributary roads	Mean
	Radial extent of ring roads	Within inland sample city range and close to mean
	Shortest distance from motorway to inner ring	Within inland sample city range and close to mean
	Max slope threshold	Within inland sample city range and close to mean
	Associations with topography	Main road network should follow the terrain
Rail network	Length of rail network	Within inland sample city range and close to mean
	Number of main lines	Mean
	Number of intersections	Mean
	Shortest distance from main line to inner ring	Within inland sample city range and close to mean
	Max slope threshold	Within inland sample city range and close to mean
	Associations with topography	Rail network should follow the terrain
Minor road network	Land cover class	Interdependency based on inland cities
	Distance to main roads	Interdependency based on inland cities
	Altitude	Interdependency based on inland cities
	Slope angle	Interdependency based on inland cities
	Distance to city centre	Interdependency based on inland cities
Land cover	Distance to main roads	Interdependency based on inland cities
	Altitude	Interdependency based on inland cities
	Slope angle	Interdependency based on inland cities
	Distance to city centre	Interdependency based on inland cities
Population	Land cover class	Interdependency based on inland cities
	Road density	Interdependency based on inland cities
	Altitude	Interdependency based on inland cities
	Slope angle	Interdependency based on inland cities
	Distance to city centre	Interdependency based on inland cities

The analyses of the urban components has further shown that the spatial distribution of the minor road network, the land cover and the population density indicate very intricate pattern that are characterised by the interdependencies with the other urban components. The analysis has also detected changing spatial patterns with distance from the city centre, contributing further to the urban complexity. These patterns have to be reflected in the formulation of the structural and design rules. The rules for the minor road network, the land cover and the population density, therefore, focus firstly, on the interaction between these urban components, secondly, on the interactions of these urban components with the topography and thirdly, on the reflection of centrality. The considered associations are outlined in Table 1.

Constructing approach: terrain and transportation network

Following the formulated rules, terrain and transport routes are constructed without additional input, purely based on the specifications of the inland sample cities. The design rules require the terrain and transportation network of SIENA to follow average values of the inland sample cities for all specified characteristics. In case of topography the minimum and maximum altitude values are predefined as is the altitude range. The main transportation network is even more confined by the rules, specifying the number, length and spatial arrangement of roads and rail lines and their interactions with the topography.

The topography for SIENA is based on an existing area of terrain, outside the sample area. Areas are sought that have an average altitude and slope similar to that of the inland sample cities, and which contain a small, old town on a river, that could be regarded as the historical focus of the simulated city. The area of Axminster is chosen, a town in Dorset, founded in the middle of the seventh century, whose topography broadly resembles that of the sample cities. The DTM for

Axminster and its surroundings is downloaded from Digimap and then clipped to an area of the appropriate dimensions (11.3 km x 13.5 km), using the town centre of Axminster as the centre. The coordinates of the resulting DTM are rescaled to create a new coordinate system for SIENA, attributing the lower left corner with x, y-coordinates of 0,0. The altitude range of the chosen terrain does not exactly match the average altitude range of the inland sample cities. Therefore, the altitude is adjusted to correspond with the average altitude values of the inland sample cities. The DTM thus generated is the constructed topography for SIENA (Figure 4).

Other possible methods to construct the terrain for SIENA include the application of a polynomial surface by averaging the polynomial functions for the sample cities and then applying the resulting polynomial function to SIENA (*Nico et al., 2005*). This, however, results in an overly smooth terrain and methods to enhance the surface by introducing random variation, by creating channels using flow routine models, and by rubber-sheeting to distort the surface do not prove successful. The decision was therefore made to construct the topography instead as described above.

The design rules require different approaches for the modelling of the different elements of the transport network. The main roads and the railway network are defined by the inland city averages and the association with the terrain (see Table 1) whilst the distribution of minor roads is determined by the spatial interactions between the minor roads and the other urban components. Distributary roads are defined here as the connection between the main road network and the minor network in areas of high population densities. Therefore, different approaches are used to model the different entities of the transport network (see Figure 3). Motivation for the choice of each modelling approach is given in the relative sections below.

To model the main roads and the rail network for SIENA an empirical approach is used in order to observe the design rules outlined in Table 1. Using a deductive method by which each road type is successively digitised allows considering the observed spatial relationship between the different road segments and road types as well as the associations with the terrain. For this purpose, the main roads in SIENA will be categorised into the same road types as identified for the sample cities: ring roads, feeder roads and motorways.

The main roads and railway in SIENA are digitised as shapefiles in ArcMap to be integrated into the GIS. Ring roads are the first road type to be modelled. They are digitised as circle like features around the city centre with numbers and diameter values close to the mean of the inland sample cities and the exact location determined by the terrain, more precisely, the slope threshold. The motorway is then digitised in relation to the position of the ring roads (mean shortest distance from inner ring road to motorway) and the topography (slope threshold) taking the mean length of motorways in the inland sample cities into account. Then the feeder roads are digitised, again following the observed spatial relationship with the already digitised road types and observing the structural rules (mean values for number of feeder roads, length of roads and the slope threshold). The relationship with the terrain is respected by digitising feeder roads along the valleys of the terrain. Railways are digitised to follow the terrain, taking the slope threshold for railways into account. The length and number of junctions is determined by the structural rules derived from the statistical analysis of the rail network (mean values based on inland sample cities).

Figure 4 shows the modelled transportation network with the topography in the background.

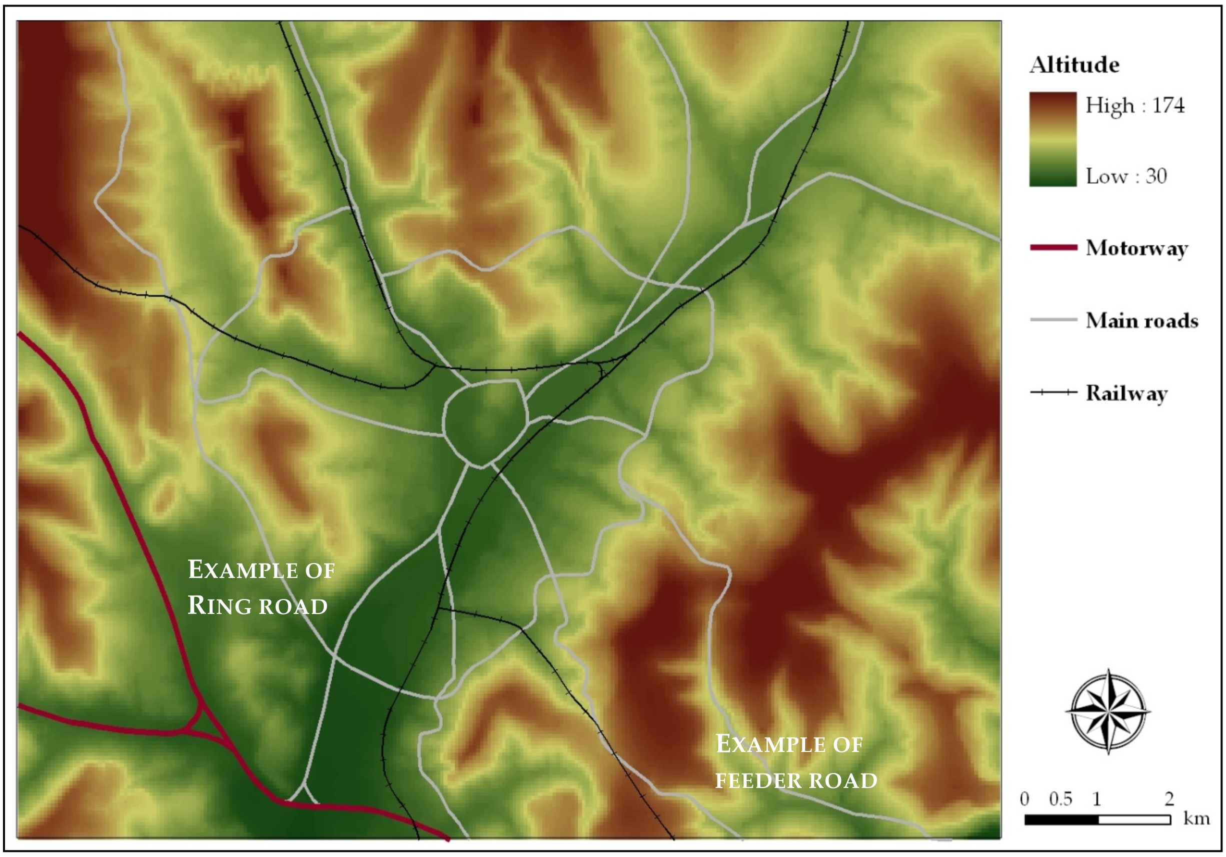


Figure 4. Topography and main transportation network for SIENA

Probabilistic modeling approach: land cover, minor roads and population densities

A probabilistic modelling approach is chosen to model the land cover, the minor road density and the population for SIENA. Probabilistic models are tools to estimate the probability of a distribution occurring again on the basis of previous distributions. The model used here is based on the assumption that the urban components have a certain probability to occur at a certain location within the urban area based on their spatial pattern and their associations with the other urban components as summarised in the design rules. The probabilistic modelling approach allows to be focused on the spatial considerations of urban structure, one of the requirements of SIENA (see above). Hence, the main focus of the probabilistic model lies on the estimation of the spatial distribution of the land cover, minor roads and population together, with their interdependencies.

Effective tools for modelling dynamic processes such as land cover change over time, the growth of a city or traffic flow (*Alperovich & Sopasakis, 2008; Parker et al., 2003; Sante et al., 2010)* vary from dynamic approaches such as the cellular automata and the multi-agent paradigm to the here applied probabilistic approach. For this study the decision is made not to organically grow the simulated urban area but to model a static city structure representing a certain point in time in order to focus on the spatial aspects of urban structure rather than temporal ones. Dynamics can be introduced later to SIENA by adding spatial-temporal contextual data onto the fixed structural foundation of the simulation or during the modelling of scenarios.

The probabilistic modelling approach applied to model land cover, minor roads and population density for SIENA uses so-called 'parameters' to calculate for each 25 x 25 m grid cell the probability of an urban component to occur as a certain type, referred to here as 'class'. A parameter is defined as a measure of class membership at the 25 x 25 m grid level for each of the specifications outlined in the design rules (see Table 1). The list of parameters used to calculate the probabilities of class membership is given in Table 2 for each urban component. Table 2 also lists the available classes for each urban component: in case of land cover, 'class' corresponds to the seven land cover classes built-up areas, continuous urban, industrial land, woodland, arable land, managed grassland and unmanaged grassland; minor road density is categorised in fifteen classes increasing in minor road density; in case of population, 'class' refers to the number of people living in each 25 x 25 m grid cell.

Table 2. Probability model specifications for the urban components

Urban component	Parameters	Class membership	
Land cover	Distance to main roads	Class	Land cover class
	Altitude	1	build-up areas
	Slope angle	2	continuous urban
	Distance to city centre	3	industrial land
		4	woodland
		5	arable land
		6	managed grassland
		7	unmanaged grassland
Minor road density	Distance to main roads	Class	Minor road density (m/m²)
	Land cover class	1	0
	Altitude	2	> 0 – 0.0015
		3	> 0.0015 – 0.0030
	Slope angle	4	> 0.0030 – 0.0045
	Distance to city centre	5	> 0.0045 – 0.0060
		6	> 0.0060 – 0.0075
		7	> 0.0075 – 0.0090
		8	> 0.0090 – 0.0105
		9	> 0.0105 – 0.0120
		10	> 0.0120 – 0.0135
		11	> 0.0135 – 0.0150
		12	> 0.0150 – 0.0165
		13	> 0.0165 – 0.0180
		14	> 0.0180 – 0.0200
		15	> 0.0200
Population	Minor road density	Class	Population
	Main road density	1	1
	Land cover class	2	2
	Altitude	3	3
	Slope angle	4	4
	Distance to city centre	...	...
		523	523

In order to calculate probabilities of class membership in SIENA, associations between the parameters and class membership in the sample cities are established. Gridded points with a 25 x 25 m resolution are, therefore, created for each inland sample city as well as SIENA. The relevant parameters such as altitude, slope angle, distance to city centre, etc. as outlined in the design rules are calculated for each gridded point. In addition, class membership for each urban component is extracted for gridded points in the inland sample cities. Probabilities of class membership are then established for each gridded point in the inland sample cities. These probabilities are based on the joint distribution of the parameters relevant for the respective urban component. Discriminant analysis is applied in SPSS to derive probabilities for class membership of each gridded point, as a function of the parameters. This function can then be applied to the unclassified cells in SIENA to assign probabilities of class membership, the so-called 'allocation probability'.

The function takes the general form:

$$\boldsymbol{d_{ik} = b_{0k} + b_{1k}x_{i1} + \cdots + b_{pk}x_{ip}} \qquad \textbf{Equation 1}$$

where, d_{ik} is the value of the k^{th} discriminant function for the i^{th} case, p is the number of parameters, b_{jk} is the value of the j^{th} coefficient of the k^{th} function, x_i is the value of the i^{th} case.

To allocate each gridded point in SIENA with a class, the cumulative probability of the allocation probability across all classes for one urban component is determined (scored from 0 to 1). To select a class for any unclassified gridded point in SIENA, a random number between 0 and 1 is drawn, and this used to allocate the matching class. The approach is summarised in Figure 5.

The land cover model

Statistical exploration of the land cover identified seven main land cover classes that are dominant in all sample cities: built-up areas, continuous urban, industrial land, woodland, arable land, managed grassland and unmanaged grassland. Other land cover classes account in total for only a small part of the total area (< 5%) and are largely unrelated to the other urban components. The land cover model is therefore restricted to the seven main land cover classes. Other, minor land classes such as dump sites can, if required, be added later to the simulation as contextual data.

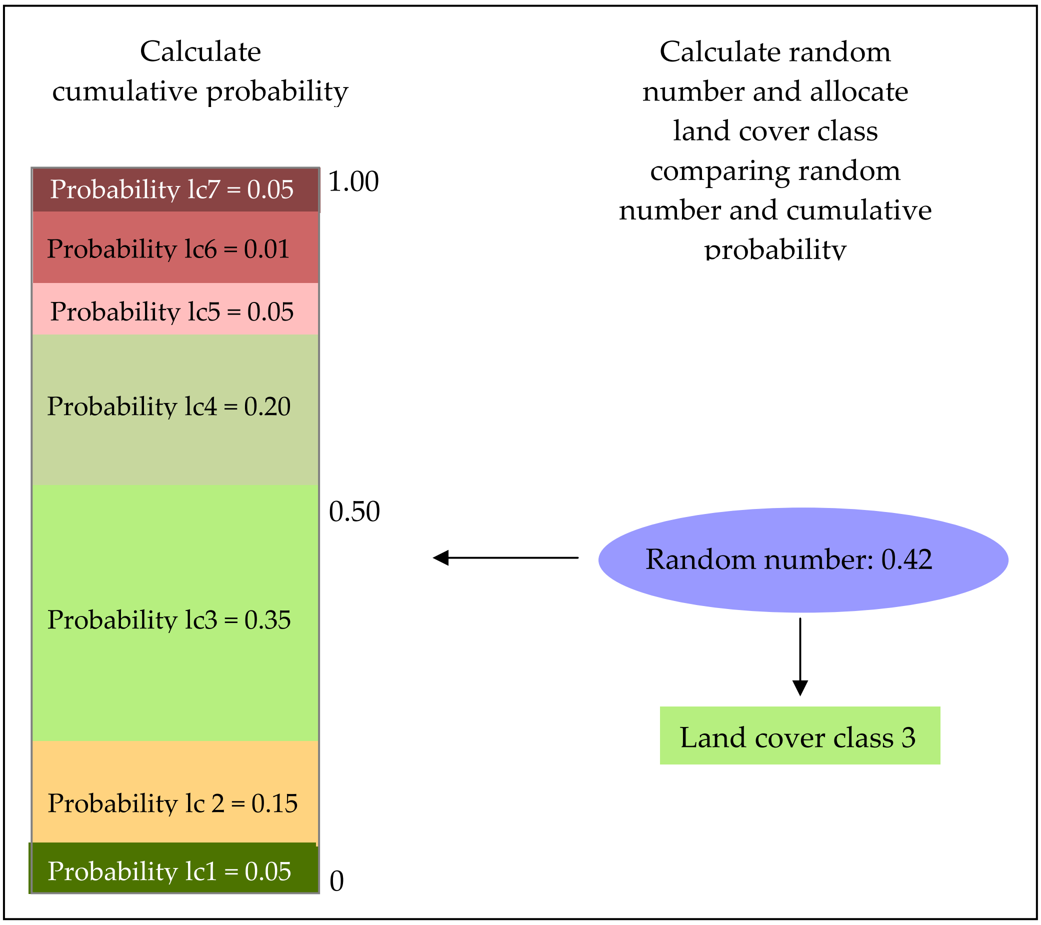


Figure 5. Example of allocating a land cover class by randomly selecting from the cumulative probability

Using the allocation probability to derive land cover class membership alone does not account for the size of the different land cover classes as seen for the inland sample cities. Instead, it generates a very granular land cover surface that over- or under- represents certain classes. To correct for differences in class size, a 'stress measure' is introduced in the land cover model that accounts for the number of grid cells attributed with a certain land cover class. 'Adjusted probabilities' for each land cover class are then recalculated as:

$$P_{ij} = P'_{ij}S_i \qquad \textbf{Equation 2}$$

where P_{ij} is the adjusted probability of land class i in cell j, P'_{ij} is the previous allocation probability for land cover class i in cell j, and S_i is the stress measure for land cover class i.

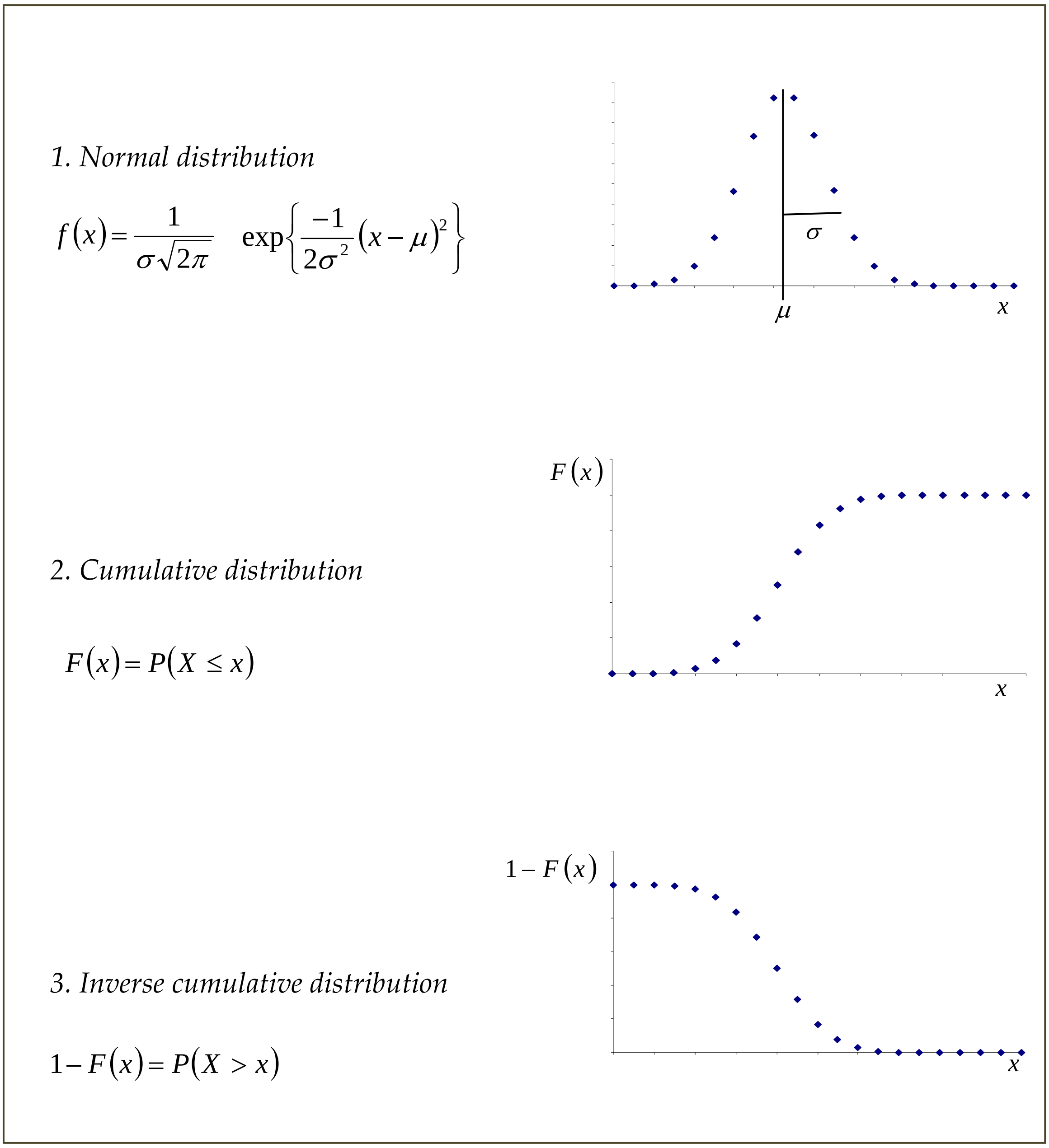


Figure 6. Calculation of the inverse cumulative distribution of class size

In order to establish the stress measure for each land cover class, an inverse cumulative distribution of class size is calculated to extract realistic target values for class size. A normal distribution is assumed for the class sizes as seen in the mean of the sample cities. The standard deviation is chosen to be half the mean to allow for a wide distribution because the standard deviation of the sample cities results in a too narrow distribution. Then the inverse cumulative probability, the stress measure, is calculated of the likelihood to find the total class size based on the normal assumption (see Figure 6).

Figure 7 shows as an example the stress measure for the land cover class continuous urban. As can be seen from the inverse cumulative probability distribution, the probability to allocate more gridded points with a certain land cover class drops rapidly after the target size for the land cover class is reached. This avoids the over-representation of certain land cover classes in the model.

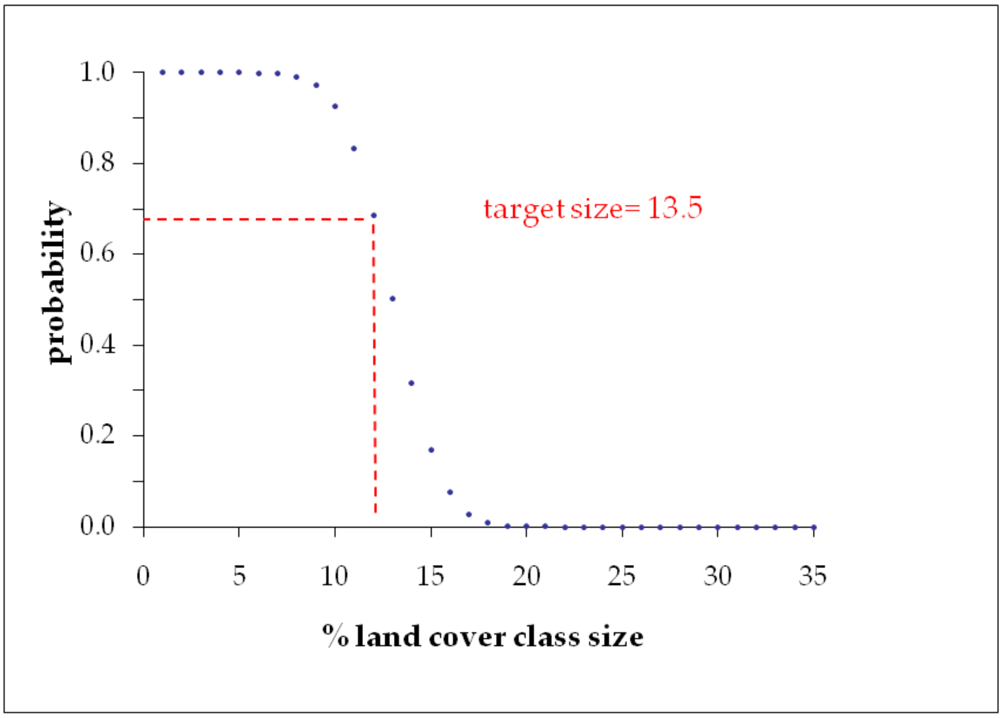


Figure 7. Distribution of cumulative probability for land cover class continuous urban

The land cover model is applied in stages. First, 10% of the 25 x 25 m gridded points are randomly selected and assigned a land cover class based on the allocation probability. This is done by visiting each gridded point in turn, selecting a random number between 0 and 1 and comparing it with the cumulative allocation probability, as shown in Figure 6. A further 10% of gridded points are then randomly selected and the adjusted probabilities of class membership are calculated by weighting the allocation probability with the stress measure and therefore taking account of the area of land already allocated to each land cover class. This process is repeated until all gridded points are attributed with a land cover class.

The resulting 25 x 25 m land cover surface shows a generally recognisable pattern of land cover class distribution. However the surface is still very granulated and scattered and needs to be enhanced in order to simulate realistic land cover patterns. To reduce granularity and ensure that grid cells cluster into realistically sized land cover patches a filter is applied to smooth the land cover surface. Tests are carried out to establish the optimal filter size for the smoothing process. This has to be done both in view of shape and size of the resulting land cover class patches as well as with regard to maintaining the established target size for land cover classes.

The smoothing of the land cover surface is done using the command FOCALMAJORITY in the GRID environment in ArcInfo. FOCALMAJORITY locates the majority value, i.e. the value that appears the most in the filter neighbourhood and attributes the centre cell of the grid with that value.

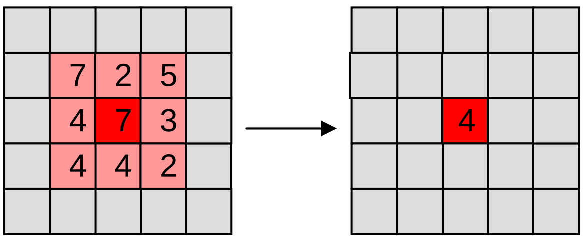


Figure 8. Example of FOCALMAJORITY applied to a 3 x 3 neighbourhood

Figure 8 shows as an example a 3 x 3 neighbourhood. Value 4 is in the majority within the neighbourhood and the centre cell is therefore attributed with this value.

To establish the optimal filter size for the smoothing of the land cover surface, land cover is modelled for the two sample cities Leicester and Reading using the same approach as for SIENA. Filters of various sizes, ranging from 3 x 3 neighbourhoods to 11 x 11 neighbourhoods, are applied to the modelled land cover classes. Comparisons are made between the modelled land cover using the different filters and the observed, real-world land cover surface, using accuracy assessment (aka confusion matrix) methods.

modelled land cover classes

observed land cover classes

# cells	lc1	lc2	lc3	total
lc1	50	5	2	57
lc2	14	13	0	27
lc3	3	5	8	16
total	67	23	10	100

$$A_T = \frac{C_C}{\sum C}$$

$$A_{Ci} = \frac{C_{Ci}}{\sum C_i}$$

A_T is the total accuracy, C the number of cells and C_C the number of correctly classified cells.
A_{Ci} is the class accuracy for the land cover class i, C_i the number of cells in land cover class i and C_{Ci} the number of correctly classified cells in land cover class i.

Figure 9. Accuracy assessment of land cover classes: diagonals represent cells classified correctly according to observed data (green); off-diagonals are misclassified cells (white)

Accuracy assessment is often used in remote sensing to evaluate the classification of remotely sensed data. Figure 9 shows an example based on the land cover model. Observed land cover classes are mapped against the modelled land cover classes. The total accuracy is then calculated by dividing the number of correctly classified cells (shown on the diagonal, in green) by the total number of cells (dark brown). In this example 71% of all cells are classified correctly.

The significance of the accuracy value is analysed by computing the Kappa statistic (*K*). *K* ranges from 0 to 1, and gives a direct measure of the probability that the accuracy value could not have occurred by chance: a Kappa result of 0.85, for example, means that there is an 85% better agreement between the observed and modelled land cover classes than by chance alone (*Congalton, 1991*).

Kappa was computed as:

$$\boldsymbol{K} = \frac{N\sum_{i=1}^{r} x_{ii} - \sum_{i=1}^{r}(x_{i+}x_{+i})}{N^2 - \sum_{i=1}^{r}(x_{i+}x_{+i})} \qquad \textbf{Equation 3}$$

where N is the total number of cells in the matrix, r is the number of rows in the matrix, x_{ii} is the number of row i and column i, x_{+i} is the total for row i, and x_{i+} is the total for column i (*Jensen, 1996*).

The total accuracy and class accuracy are computed for each filter size, using the ArcView extension Kappa Analysis 2.0, and the Kappa statistic is then calculated (*Jenness & Wynne, 2005*). Results for the different filter sizes are then compared in terms of Kappa. The results suggest that the application of two successive filters, using a 3 x 3 neighbourhood, provide the best predictions of reality. These filters are therefore applied to the modelled 25 x 25 m land cover surface of SIENA. The resulting land cover is shown in Figure 10.

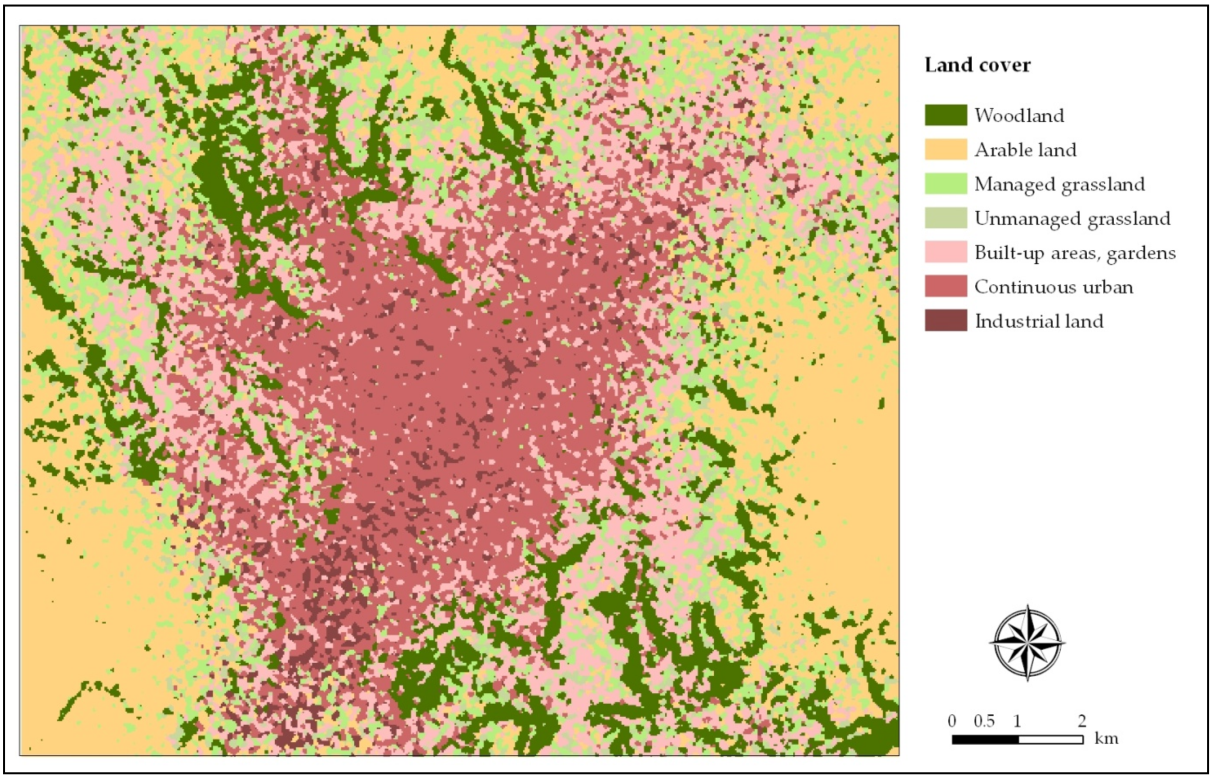


Figure 10. Modelled land cover of SIENA

The minor road model

Minor roads are essential features in the urban environment as they connect point locations such as houses in residential areas, workplaces or shops to the main road network. This information is, for example, important when modelling daily activity patterns or journeys of individuals for personal exposure assessments. Further, minor roads are line sources of emissions and contribute to the total traffic emissions within an urban area.

Minor roads are modelled in SIENA as minor road densities for each grid cell. This allows conclusions about the total minor road length within an area. Minor roads are too numerous and their ramifications too complex in terms of connectivity to be realistically constructed as line features.

The minor road model is based on the same probabilistic modelling approach as described for the land cover model using the parameters outlined in Table 2. But a 25 x 25 m resolution proves to be too small to detect any variation in minor road density distribution between the grid cells within the sample cities because most cells will have a minor road density of zero or close to zero. The resolution of the grid, therefore, has to be increased. To identify the optimal resolution for the minor road model the minor road density distribution is analysed for different spaced grids. Again, one of the sample cities, Leicester, is used as a test bed. Different grids with resolutions of 25 x 25 m, 50 x 50 m, 100 x 100 m, 250 x 250 m, 500 x 500 m and 1000 x 1000 m are created and the minor road density calculated for each grid cell. The best results are achieved with a 250 x 250 m resolution grid because here the minor road densities provide enough variation between the grid cells to make an informed decision about the spatial distribution of minor road densities and, on the other hand, the resolution is still fine enough to provide enough detail for the purpose of SIENA. The parameters altitude, slope angle, land cover class, distance to city centre and distance to main roads are, therefore, calculated for the inland sample cities for each 250 x 250 m grid cell. In case of altitude the average altitude for each 250 x 250 grid cell is calculate, in case of land cover, the dominant land cover class is attributed to the 250 x 250 m grid cell.

As outlined in Table 2, the minor road density is grouped into fifteen minor road density classes. This allows a wide variation of classes with approximately equal numbers of grid cells in each class. Discriminant analysis is run to derive the allocation probability of class membership and minor road classes are then attributed to each 250 x 250 metres grid cell based on the cumulative probabilities. To avoid overrepresentation of some density classes the class membership is, again, confined based on the stress measure previously described for the land cover model.

The population model

The population model differs slightly from the land cover and minor road model. The spatial analysis of the population distribution within the sample cities has shown that there are very strong associations between the land cover and the population and a distinguishing population density is identified for each land cover class. The population is, therefore, first calculated at land cover patch level by applying the average population density of the inland sample cities for each land cover class to the land cover patches of the urban simulation as follows:

$$\boldsymbol{P_{ij} = \frac{A_{ij}}{D_i}} \qquad \textbf{Equation 4}$$

where P_{ij} is the population for land cover patch *j* in land cover class *i*, A_{ij} is the area of the land cover patch *j* in land cover class *i* and D_i the population density for land cover class *i* derived from the inland sample cities.

The population at patch level is further disaggregated to the 25 x 25 m grid level using the probabilistic modelling approach. Again, discriminant analysis is used to derive the allocation probability for class membership based on the remaining parameters specified in Table 2. Each 25 x 25 m grid cell is then attributed with the number of people, accordingly (see Figure 11). Grid centroids of cells, which have a population greater than zero are defined as a residential address location.

Spatial analysis approach: distributary roads

Apart from the main and minor roads, a third road type is modelled for SIENA, so-called 'distributary roads'. After a visual exploration of the mapped main and minor roads in conjunction with the population distribution, the need for a third road type becomes apparent that provides a link between residential areas and main roads. Distributary roads are defined in SIENA as roads that provide connections between the main road network and the minor roads in areas of high population density. They are situated between main roads and minor roads in the road hierarchy.

To model the distributary roads spatial analysis tools available in GIS are used (*Longley & Batty, 1996*). Tools exist, for example, for measuring distance between two locations, both in Euclidean distance and, in terms of factors such as road infrastructure or travel cost. The spatial analysis approach makes use of some of these tools to model the distributary roads for SIENA.

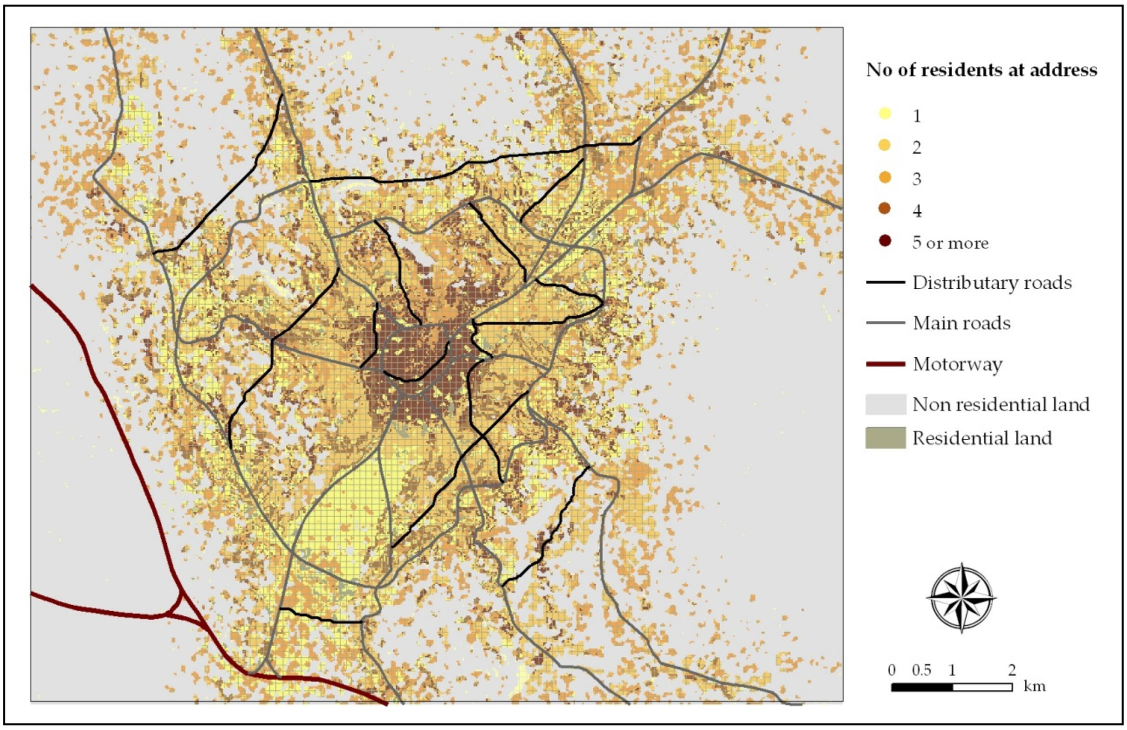


Figure 11. Modelled population and distributary roads

The COSTPATH command in ArcInfo provides a tool to calculate the least-cost path that follows the route of the highest population density based on an origin and destination, in this case the main roads. A cost-weighted distance function modifies the Euclidian distance by equating distance with the cost of travel (*ESRI, 2001*). The shortest path is then calculated from a source to a destination over the cost surface. The ‘cost’, in this case, is the population density. Two 25 x 25 metres grids are created identifying the start and end points for the distributary roads. Start and end points are randomly located close to feeder and ring roads within areas of high population densities. The number of distributary roads is defined by the design rules. Then the grid holding the population information is created which is used as the cost for the least-cost path. The COSTPATH command then identifies routes from the start point to the end point following the path of the highest population density. The COSTPATH command is run 20 times for each distributary roads to obtain an array of possible least-cost paths. Based on the suggested routes, line features are digitised reflecting the line of best-fit. These lines represent the distributary roads for the SIENA and are incorporated in the already simulated road network (see Figure 11).

Model evaluation: results and quality assessment

A quality assessment is carried out to evaluate the representativeness of SIENA to real-world cities. This allows making assumptions about how well the core data structure of SIENA imitates the data structure of the inland sample cities. For this purpose, measures calculated for the urban components of the inland sample cities in the statistical analysis are applied to SIENA. Statistics are then computed to assess the comparability of the simulation results to the inland sample city results.

To assess the constructed topography, one-sample T-tests are carried out for mean altitude and slope angle values for each of the four zones increasing in distance from the city centre. The T-test assesses the difference between the inland sample city mean values and the values of SIENA. During the construction process, the simulated topography has been confined to the mean altitude values of the sample cities and minimum and maximum values therefore do not need to be assessed. The results of the T-tests suggest that the simulated topography imitates the mean topography values of the inland sample cities with p-values > 0.05 and 90% confidence interval. This is an indication that the selection of Axminster as a proxy topography for SIENA has been a suitable choice.

The main transportation network is also tested with the one-sample T-test. The measures used for the spatial exploration are calculated for the transportation network of SIENA and then compared to the mean values of the inland sample cities. Again, the T-test cannot detect any significant difference between the inland sample city mean and the SIENA value. This is also true for the minor road density by distance from the city centre. The only exception is the percentage area of the different land cover classes next to the motorway. This can probably be explained with the location of the motorway. In some sample cities the motorway reached far into the city and ends close to the city centre, whilst in others the motorway runs along the periphery, as is the case in SIENA. Depending on the location of the motorway different dominant land cover classes can be found next to the motorway.

The quality assessment for the land cover is split into the three different hierarchical levels, city level, land cover class level and land cover patch level, as used for the statistical exploration of the inland sample cities. At the city level, one-sample T-test results detect differences between sample city means and SIENA values. As would be expected, the city size is comparable to the inland sample city mean, however, the number of patches, the land cover class richness as well as the Shannon's Diversity and Shannon's Evenness Index differ. These differences can be explained by the land cover model. The number of land cover classes, the richness, is constrained in the model to the seven main land cover classes while the inland sample city consist of main as well as minor land cover classes. This difference in the number of land cover classes also explains the differences detected for the Shannon's Diversity and Shannon's Evenness Index. The significant difference in the number of patches between the inland sample city mean and the urban simulation value is a result of the model premise to model the land cover at the 25 x 25 m grid resolution and probably also an effect of the applied smoothing filter.

This pattern is repeated for the quality assessment at land cover class level. Measures of location such as the adjacency measure and the topographic influence on the land cover class distribution are comparable whilst measures relating to land cover class size show, especially for the urban land cover classes, significant differences between the sample city mean and the SIENA value.

Nonparametric tests are used for the analysis at land cover patch level because, as the exploration of the sample cities has shown, the measures at land cover patch level are extremely skewed. In this case, the T-test does not produce any conclusive results because it assumes normally distributed variables. Nonparametric tests, on the other hand, do not make assumptions about the parameters of a distribution and, similar to the T-test, help determining if values of a particular variable differ between two groups. Nonparametric tests chosen for the quality assessment are the Mann-Whitney U test and the Kolmogorov-Smirnov Z test. The Mann-Whitney U test assesses the null hypothesis that two independent samples come from the same population. The Kolmogorov-Smirnov Z test is a goodness-of-fit test which tests the null hypothesis that two distributions are not significantly different. The Kolmogorov-Smirnov Z test, however, is sensitive to differences in location and shape. Centering each variable around its mean eliminates differences and allows for comparison of the distribution shapes. If the SIENA values are compared to the values of the sample cities the statistical power to detect significant differences between the two groups is very high due to the high number of records. To get a more differentiated picture, the SIENA values are therefore compared to each city and land cover class separately. For this analysis, patches smaller than 0.5 ha are excluded from the inland sample cities because they are a product of the data manipulation as already discussed in the previous chapter. Patches smaller than 0.5 ha in SIENA are, however, included because they are a result of the land cover model at 25 x 25 m grid resolution. The statistical tests do not detect any significant differences between SIENA and the inland sample cities in terms of the spatial distribution of the land cover patches. But measures of patch size and patch form such as circularity do differ for all land cover classes and sample cities.

To assess the population model the one-sample T-test is again used. Overall the mean values of SIENA compared to the inland sample cities suggest that the simulated population displays similar spatial patterns and densities than the inland sample cities. The only exception is population density within 1 km of the city centre which the population model over-predicts. This is probably due to too much weight being put on the centrality component in the population model.

Overall the quality assessment confirms that the core structure of SIENA is similar to the structural elements in the inland sample cities and especially in terms the interaction between the elements and their spatial distribution. Scenarios and data modelled using SIENA are not directly transferable to the inland sample cities or other real-world cities but they allow us to draw conclusions about real-world situations and processes. No absolute results can be obtained but general tendencies or trends can be provided using SIENA.

3.2 Contextual data

Contextual data is real-world data that is modelled onto the existing core data structure of SIENA. It can be added for specific applications to provide a sound, scenario specific information structure. The nature of this data depends on the requirements of the scenario. Contextual data could be anything, from a single value to characterise the urban simulation, for example, annual mean temperature, attribute data for the urban components, such as socio-economic characteristics of the population or traffic densities of the roads, to new spatial entities such as monitoring sites. The methods to model the contextual data depend on its nature and its intended purpose.

Two approaches are introduced here. Firstly, real-world data is added to SIENA as it is, without any spatial considerations. This approach will be demonstrated with the introduction of meteorological conditions and housing characteristics. Secondly, contextual data can be added to the core data using the spatial distribution of the data in the sample cities and then applying these spatial patterns to SIENA. This approach will be shown using the example of modelling traffic volume, point emission sources and a fixed air pollution monitoring network. The contextual data described here is modelled for SIENA with view of the data requirements of the case studies carried out in the third part of the thesis.

Meteorological data

Knowledge about meteorological conditions within an urban area provide valuable information for epidemiologists both as exposure factor such as heat exposure (*Rey et al., 2009)* and as factor influencing exposure to, for example, air pollution (*Ghazali et al., 2010*). Different meteorological scenarios are modelled for SIENA making use of real-world meteorological data from Bristol, one of the sample cities. Information on wind speed, wind direction, cloud cover and air temperature are downloaded from the British Atmospheric Data Centre website. Four meteorological scenarios are selected based on these hourly data to simulate the four seasons. Wind speed and temperature together with precipitation are the most important factors characterising the different seasons (*Brunt, 2007*). If the temperature and wind speed (hourly data averaged per day) are plotted against each other, different meteorological scenarios can be distinguished. Four 14-day periods are chosen from the annual daily mean data to represent a cold and calm period (winter scenario), a warm and windy period (spring scenario), a hot and calm period (summer scenario) and a cool and windy period (autumn scenario).

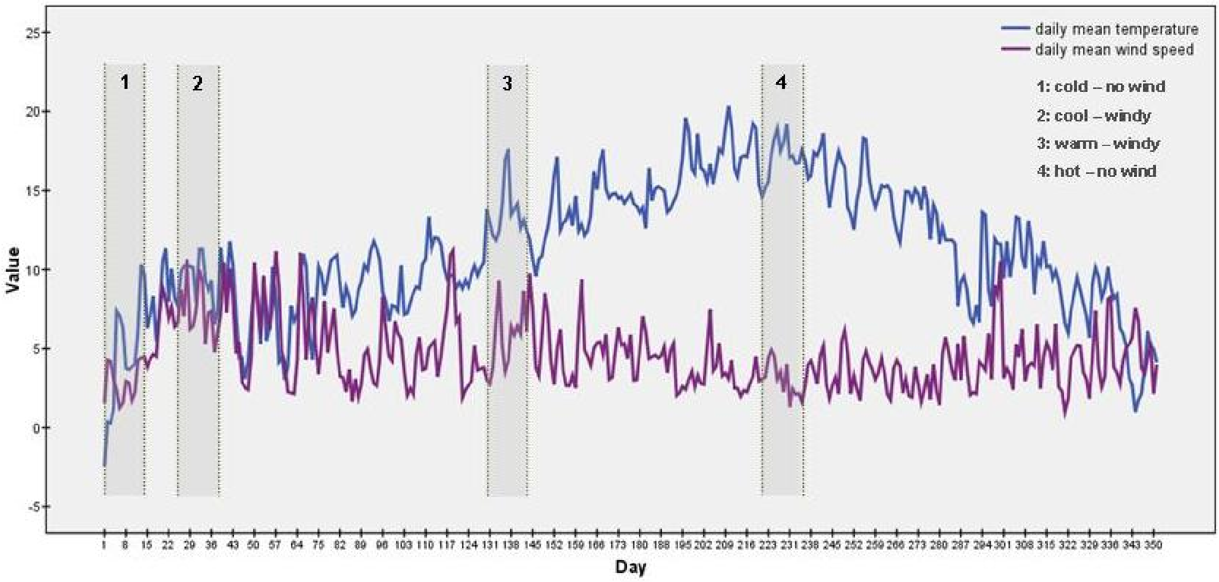


Figure 12. Meteorological data for Bristol with periods chosen to represent seasonal scenarios highlighted

As Figure 12 shows, data for the four season scenarios are not necessarily drawn from the respective season in the real-world. The approach taken to simulate the different meteorological scenarios, however, allows doing this. Real-world data is simply taken as input to put SIENA into a certain context, in this case a meteorological context. The data is added without any modification and no spatial or temporal considerations are taken into account.

Housing characteristics

Given that people tend to spend most of their time indoors, exposure to indoor air pollution is increasingly analysed in exposure assessments (*de Hartog et al., 2010)*. Besides indoor pollution sources such as gas cookers and tobacco smoke, outdoor air pollution entering the building is one of the main pollution sources of indoor air (*Mitchell et al., 2007*). Housing characteristics play here an important role in determining the amount of outdoor air entering the house by influencing the ventilation and penetration properties. Various housing characteristics are often key variables in models quantifying indoor air pollution based on outdoor air (*Dimitroulopoulou et al., 2006; Nazaroff, 2004; Wallace, 1996*). Within SIENA housing characteristics are modelled for each residential address as outlined in Table 3.

Table 3. Housing characteristics and calculation methods

Housing characteristic	Calculation
Building height: height of window in living room	Each living room window is randomly attributed with a height as follows: 40% defined as1st floor room: 1m 30% defined as 2nd floor room: 4m 10% defined as 3rd floor room: 7m 10%defined as 4th floor room: 10 m 10% defined as 5th floor room: 13 m
Room height: distance from ceiling to floor	Each room is attributed with a random number between 2.5 m and 3 m
Room area: floor area of living room	Each room is attributed with a random number between 10 m^2 and 30 m^2
Wall orientation: compass direction of external living room wall	A random azimuth value between -180 and 180 is computed for each external living room wall
Inlet area: area of controlled inlets such as doors and windows	$la_{(i,t)} = RAN \times Wa_i$ $Wa_i = Rh \times \frac{Ra}{2}$ where $la_{(i,t)}$ is the inlet area, RAN a random number between 0.1 and 0.4 assuming that inlets cover between 10% and 40% of the external living room wall Wa_i, Rh is the room height and Ra the room area

The variables modelled to define the housing characteristics correspond with the input variables for the INDEX model (*Jamieson, 2010*) which will be used to model indoor air pollution for the living room of all houses. The location of a house is assumed to be the residential address.

Similar to the meteorological conditions, this example shows how contextual data can be added to the urban simulation using real-world data or, in this case, data based on real-world statistics and simply attribute it to SIENA. Again, no spatial considerations are taken into account. In the next section, methods are introduced that first explore the spatial distribution of variables in the real-world and then apply these to model contextual data for SIENA.

Traffic volume

Information about traffic volume has been used in the past to provide environmental epidemiological studies with an indicator for exposure to traffic related air pollution (*Carr et al., 2002; Gauderman et al., 2005)*. Data on traffic volume is further used in exposure studies as input data to model emissions from traffic sources (*Goyal & Mittal, 2004; Wang et al., 2010*). In order to derive traffic emissions for SIENA, traffic volume is modelled for all roads based on real-world data from Bristol. In the case of motorways, traffic volumes are sourced from the Traffic Flow Data System (TRADS) within the Highway Agency Traffic Information System (HATRIS). Two-way flows are added to give a combined flow. Data for main roads is assembled from daily traffic counts provided by the Bristol City Council Manual Count Data for 2005.

A probabilistic approach is used to attribute the motorways and main roads in SIENA with the real-world data. Traffic flows for motorways and vehicle counts for main roads are spatially analysed for Bristol focusing on road centrality and the different road types such as feeder roads, ring roads and distributary roads.

Figure 13. Modelled traffic volume: Daily traffic counts for motorways and main roads and daily vehicle kilometres for minor roads

Centrality is analysed for zones ranging from 0 – 1 km, 1 – 2.5 km, 2.5 – 5 km and > 5 km. Probability distributions of traffic flows or vehicle counts are calculated for each road type and each zone and roads in SIENA are attributed with the real-world data based on these probabilities. Traffic flow and vehicle counts are then mapped to check for continuity in the data and to avoid jumps between the different road segments (see Figure 13).

A different approach is applied to model the traffic volume for the minor roads. The spatial distribution of real-world traffic volume on minor roads is not analysed because minor roads in SIENA are not represented as line features but as an area measure of minor road density. Instead, the traffic volume is purely based on the distribution of the simulated minor road density in each 250 x 250 m grid cell.

According to the Department for Transport, the assumed average traffic flow for urban minor roads in Great Britain is 2200 vehicles per day, made up of 95% light vehicles, 3.4% heavy vehicles such as buses, coaches and heavy good vehicles and 1.6% motorcycles (*Department for Transport, 2006*). This knowledge allows computing daily vehicle kilometres for each vehicle type and 250 x 250 m grid cell as follows:

$$DVK_{ij} = d_j \times a_j \times V_i$$ **Equation 5**

where DVK_{ij} is the daily vehicle kilometres for vehicle type i in grid cell j, d_j is the minor road density in grid cell j expressed as km/km^2, a_j is the total area of grid cell j expressed as km^2, and V_i the number of vehicles per day for vehicle type i derived from the Department for Transport statistics.

Point emission sources

Point emission sources are a main contributor to urban air pollution. Depending on the source activity, emissions from point sources have important health implications as has been proven for waste management sites and industrial sites (*de Marco et al., 2010; Elliott et al., 2001; Krajcovicova & Eschenroeder, 2007; Zambon et al., 2007)*.

In order to model point emission sources for SIENA the point emissions sources in the inland sample cities are explored in terms of their activity, pollutants emitted, spatial patterns and interactions with other urban components. Information about the number, location and characteristics of point emission sources in the sample cities are extracted from the European Pollutant Emissions Register (EPER). Reflecting the average number of point emission sources in the sample cities, three point sources are modelled for SIENA. Their character and location is probabilistically determined based on processor activities and spatial patterns in the inland cities.

The most common point source activity in the inland sample citied is waste management (n=8) followed by energy industry (n=2), production and processing of metal (n = 2) and others (n=2). Based on these probabilities, two modelled processors represent waste management facilities and one energy industry facilities. The spatial exploration of the point emissions sources finds that all energy industries can be found in industrial land close to motorways and major roads. The majority of waste management facilities can be found in built-up land close to the city centre (62.5%) followed by sites in industrial land (37.5%) further away from the centre. Based on these spatial specifications, potential locations within SIENA are identified and sites randomly chosen to digitise the point location of the modelled point emissions source. After the location of the three facilities is determined, they are attributed with emissions characteristics such as exit velocity, exit temperature, stack height etc. Detailed information are, however, only available for sites in Bristol, but emission characteristics are relatively comparable between sites of the same activity. Therefore, characteristics and emissions from the Bristol A processors can be distributed to the modelled point emission sources based on activity and locations.

Routine monitoring network

Routine monitoring networks are a regulatory means to observe and assess the air quality levels in an urban area. Routine monitoring sites typically consist of continuous and automatic monitoring sites that operate often over many years and measure a range of pollutant concentrations on an hourly basis, usually including particles, NO_2, SO_2, CO_2 and ozone. Because of their high temporal resolution and their historical data coverage they are frequently used in epidemiological studies for exposure assessments, either by directly assigning people with concentrations measured at their nearest monitoring station or as input data for exposure modelling (*Briggs et al., 2010; Dockery et al., 1993*).

In order to establish a realistic routine monitoring network for SIENA, the spatial distribution of routine monitoring networks in the sample cities is explored. Both routinely monitoring PM_{10} and NO_2 sites are considered because PM_{10} monitors tend to be very sparse. Information about routine monitoring sites in the sample cities is provided in case of PM_{10} by the Automated Urban and Rural Network managed by Bureau Veritas. The NO_2 diffusion tube network is a cooperative between DEFRA and the Local Authorities to map spatial and temporal trends in NO_2 concentrations throughout the UK. The program is coordinated on behalf of DEFRA by Netcen. In both cases, information can be obtained from UK Air Quality Archive for all sites currently measuring PM_{10} and NO_2 within the boundaries of the sample cities. All sample cities have just one routine PM_{10} monitor within their boundaries, located close to the city centre. Whereas, there are on average 10 routinely measuring NO_2 sites per sample city. In order to represent a best-case scenario for SIENA, both PM_{10} and NO_2 monitoring sites together are employed to establish a routine monitoring network, which will be assumed to measure PM_{10}.

The location of the monitoring sites for SIENA is informed by the position of the fixed monitoring sites in the sample cities. Distance to transport routes as an indicator for roadside or background site, the distance to the city centre, the site environment (land cover surrounding the site) and the site location in relation to pollution concentrations are explored for the monitoring sites in the inland sample cities. Following the same approach taken to construct the core data, sample city means are again used to derive the number of roadside and background sites to be modelled for SIENA. Associations between the site location and distance from the city centre as well as the density of sites in relation to air pollution concentration are taken into account in the model. This approach results in eleven monitoring sites –

six roadside sites and five background sites. Figure 14 shows the routine monitoring network in relation to major transport links.

These examples of contextual data demonstrate how real-world information can provide scenario specific context information for SIENA, either, by adding as it is in case of meteorological conditions and housing characteristics, or by adding attributes to the existing core data structure in case of traffic volume, or by modelling new entities such as point emission sources and monitoring stations. Spatial relationships in the real-world are maintained in SIENA by applying the methods introduced for the exploration of the sample cities the construction of the core data. Spatial patterns such as centrality or interactions with transportation network and land cover are analysed in a real-world setting and these then applied to SIENA by using a probabilistic approach to model the contextual data.

In order to carry out scenarios in an environmental epidemiological context additional contextual data are imaginable. Socio-economic characteristics could, for example, be modelled onto the core population of SIENA based on the spatial distribution of socio-economic characteristic in the sample cities. Health data could be added in a similar manner depending on the scenario carried out. As already mentioned the approach chosen to model the contextual data depends on the nature of the variables. SIENA together with spatial analysis tools to analyse spatial patterns in an urban area provide the means and flexibility to add many contextual variables. The information and data structure of SIENA is therefore improved with each application.

3.3 Derived data

A further data type that completes the data structure of SIENA is derived data (see Figure 1). Derived data is modelled within SIENA based on the core and contextual data foundation. SIENA provides, thus, the data infrastructure not only to apply established models but also to test new models. Both approaches are demonstrated here. Existing models are used to derive road and industrial emission and based on these air pollution concentrations for SIENA. Administrative areas are modelled based on the population densities of SIENA. A newly developed model is finally tested to model indoor air pollution.

Emissions

Anthropogenic emissions especially from urban traffic and industrial sources have a major effect on urban air quality. Especially road-based vehicles emit large amounts of a variety of pollutants including carbon monoxide (CO), oxides of nitrogen (NO_x), volatile organic compounds (VOCs) and particulates (PM_{10}). Together with emitted pollutants from industrial and, to a lesser extent, domestic combustion processes such as sulphur dioxide (SO_2) and nitric oxides (NO) they are the main contributors to air pollution in urban centres (*Netcen, 2009*).

A bottom-up approach is used to model emissions from road and industrial sources for SIENA. Contextual characteristics of roads and point emission sources, as described in the previous section, provide the input to estimate NO_x and PM_{10} emissions. Emission rates are derived using a commercially available atmospheric Emissions Inventory Toolkit (EMIT version 2.3) developed by Cambridge Environmental Research Consultants (*CERC, 2008*). EMIT estimates emission rates for various sources based on up-to-date emission factors. Activity data such as traffic flows, fleet composition, number of vehicle kilometres etc. can be used in combination with appropriate emission factor datasets to estimates emissions when explicit rates are not available.

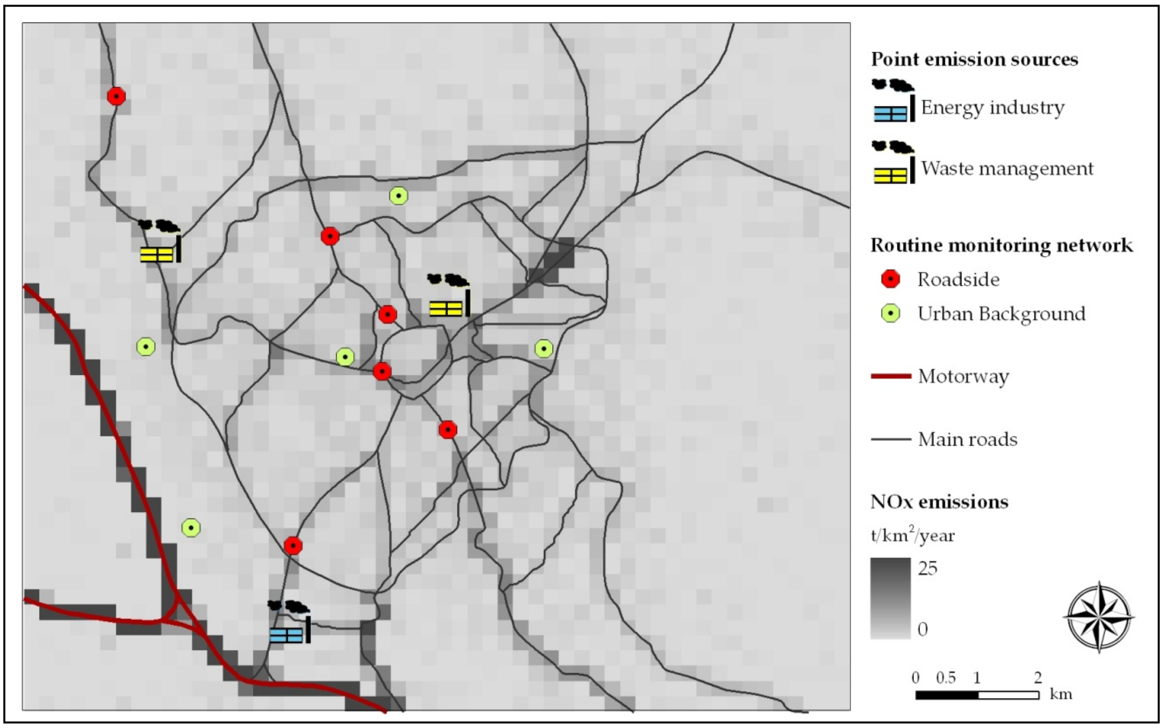


Figure 14. Total NO_x emissions from road sources including point locations of industry emission sources and the fixed monitoring network

Emissions from both transport and industrial sources are calculated with EMIT by using activity data and a pre-defined emission factor dataset held within the EMIT database. EMIT differentiates between source types. Transport sources are treated both as line sources in case of motorways and main roads as well as area sources, in case of minor roads. Industrial sites are treated as point sources. The characteristics of the different emission sources added as contextual information in the previous section are used as activity data. Emissions from minor roads are modelled within EMIT for each 250 x 250 m grid cell based on the annual vehicle kilometres, the fleet composition and road length within each grid cell. Traffic counts, fleet composition and road length are used as activity data to estimate transport emissions from line sources. Emissions from industrial point sources are entered manually based on emission factors provided for Part A sources in Bristol.

Figure 14 shows as an example modelled annual NO_x emissions from traffic sources including both line and area sources, shown here at 250 x 250 m grid resolution. As can be seen, emissions from main roads especially the motorway in SIENA dominate but minor roads emissions at the 250 x 250 m grid level can also be differentiated.

Concentrations

Once the pollutants have been emitted from a source they are dispersed in the atmosphere, a process mostly influenced by meteorological conditions. They immediately become subject to environmental processes such as diffusion, dilution, chemical reactions with other substances, transformation or decay. The amount of pollutants in the air can be measured as concentrations, typically as micrograms per cubic metre air ($\mu g/m^3$) or part per million (ppm). High concentrations of air pollutants are a major concern to human health in urban areas. Observed health effects depend on the pollutant but high concentrations are generally associated with respiratory conditions, cardiovascular diseases and even mortality (*Brunekreef & Holgate, 2002*).

Air pollution concentrations within SIENA are modelled using the Atmospheric Dispersion Modelling System (ADMS) and more specifically the multi-source dispersion model ADMS-Urban (version 2.2). ADMS-Urban, also developed by CERC, models dispersion in the atmosphere of pollutants released from industrial, domestic and road traffic sources in urban areas (*CERC, 2006*). The model makes use of point, line, area, volume and grid sources. It is designed to allow consideration of complex dispersion problems such as multiple industrial and road traffic emissions over a large urban area. ADMS-Urban applies up-to-date physics using parameterisation of the boundary layer structure based on Monin-Obukhov length and the boundary layer height which allows for a realistic representation of the changing characteristics of dispersion with height. The result is a soundly based prediction of pollutant concentrations (*Mchugh et al., 1997*).

Emission data from the traffic and industrial sources is loaded into ADMS-Urban via an emissions inventory. Surface roughness and Monin-Obukhov length are adjusted to reflect the conditions in an urban setting. The Monin-Obukhov length measures the stability of the atmosphere which in urban areas is prevented from becoming very stable by the urban heat island. In order to integrate the terrain into the dispersion model, information about the topography of SIENA is fed into the model. Although, the terrain is available at a 25 m resolution, the altitude is averaged to a 450 m resolution in order to observe the ADMS-Urban limit to the number of grid cells for the terrain file.

In the first instance, two-week average concentrations are modelled for the four different meteorological scenarios set out in the previous section. Local concentrations are computed for each meteorological scenario at a 25 x 25 m grid resolution for nitrogen dioxide (NO_2 $\mu g/m^3$) and particulate matter (PM_{10}, $\mu g/m^3$). To determine the total concentrations within the urban area a constant background concentration is added to allow for long-distance transport of air particles. Background concentrations are derived from the UK air quality report and a constant concentration of 20 $\mu g/m^3$ added for No_2 and 17 $\mu g/m^3$ for PM_{10} concentrations (*Netcen, 2009*).

Hourly concentrations are then computed based on the two-week averages using hourly traffic count data to weight the concentrations. Hourly traffic counts are not available at the road level and could, therefore, not have been used to directly inform the ADMS model. This would also have been computationally too intensive. The Department of Transport, however, produces statistics on average hourly traffic counts. Their road statistics report on traffic, speed and congestion includes statistics on traffic distribution by time of day and day of the week

(*Department for Transport, 2006*). These statistics are used to adjust the pollution surfaces by adding or removing concentrations form the original ADMS output proportionally. PM_{10} concentrations in urban areas are only partly due to traffic sources and background contribution plays a big part. The scaling of the concentration is, therefore, only done on the traffic part of the contribution. The specialist group CAFE of the European Environment Agency quantifies the traffic contribution to urban PM_{10} as 62% on weekdays and 54% on weekends (*van Aalst, 2003)*. The air pollution surface is adjusted accordingly and the weighted hourly traffic contribution added to the background part. In addition, random variation between 0% and 20% of the hourly concentrations is either added or subtracted from each 25 x 25 m grid cell as follows to allow for natural fluctuations in the concentrations:

$$Cn_{ik} = \begin{cases} C_i + \left(\frac{C_i}{100}x_i\right) & k = 0 \\ C_i - \left(\frac{C_i}{100}x_i\right) & k = 1 \end{cases} \qquad \textbf{Equation 6}$$

where Cn_{ik} is the air pollution concentration with added variation for grid cell i and condition k, C_i is the modelled concentration for grid cell i, x_i is a random value from a uniform distribution between 0 and 20 and k is a randomly attributed value of either 0 or 1. This results in 24 hourly concentration maps of NO_2 and PM_{10} for both weekdays and weekends.

The air pollution model cannot be validated because the real situation is unknown or rather does not exist. One way, however, to assess the modelled air pollution concentrations for SIENA is to compare the modelled concentrations to real-world concentrations. Hourly NO_2 and PM_{10} concentrations measured at monitoring stations within the sample cities are, therefore, compared to hourly concentrations measured at the monitoring stations in SIENA. In order to simulate the measurements obtained from monitoring stations in SIENA, the modelled hourly concentrations surfaces for NO_2 and PM_{10} are intersected with the SIENA monitoring sites and the attributed concentrations treated as the measured concentration values.

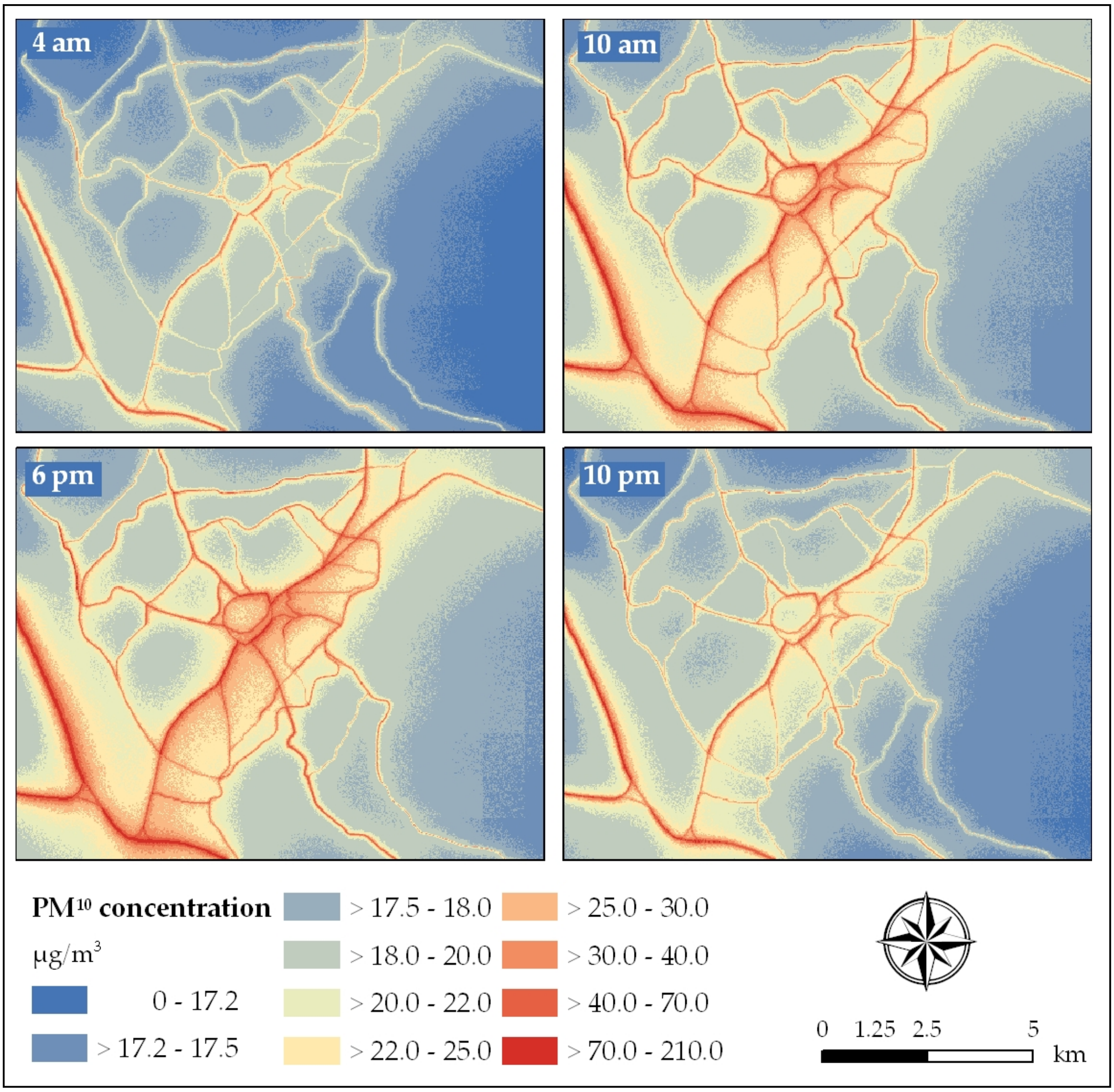


Figure 15. Four examples of hourly PM_{10} concentration maps for a weekday modelled for the urban simulation using the meteorological summer scenario

An ANOVA is carried out to compare the means of the measured concentrations in the sample cities to the measured concentrations in SIENA. The results show a similar picture for both NO_2 and PM_{10}. Hourly concentrations in SIENA are not significantly different from hourly concentrations measured in the inland sample cities. The only exceptions are night-time concentrations in Nottingham and Derby which are slightly higher than in SIENA. Overall the analysis demonstrates that the daily temporal changes in air pollution concentrations in SIENA behave comparably to real-world concentrations and the modelled concentrations are within the range of real-world concentrations.

The spatial distributions of NO_2 and PM_{10} concentrations in SIENA are further assessed visually. Figure 15 shows as an example the mapped PM_{10} concentrations for four hours on a weekday modelled with the meteorological summer scenario. As can be seen, the spatial distribution of the concentrations is strongly influenced and confined by the road network. This pattern has also been observed in real-world cities. Hoek *et al.* documented a low spatial variability away from roads for fine particles (*Hoek et al., 2002)*. A study in Basel reported that the annual average PM_{10} concentrations at a traffic site were 40% increased compared to an urban background site (*Roosli et al., 2001*).

Administrative areas

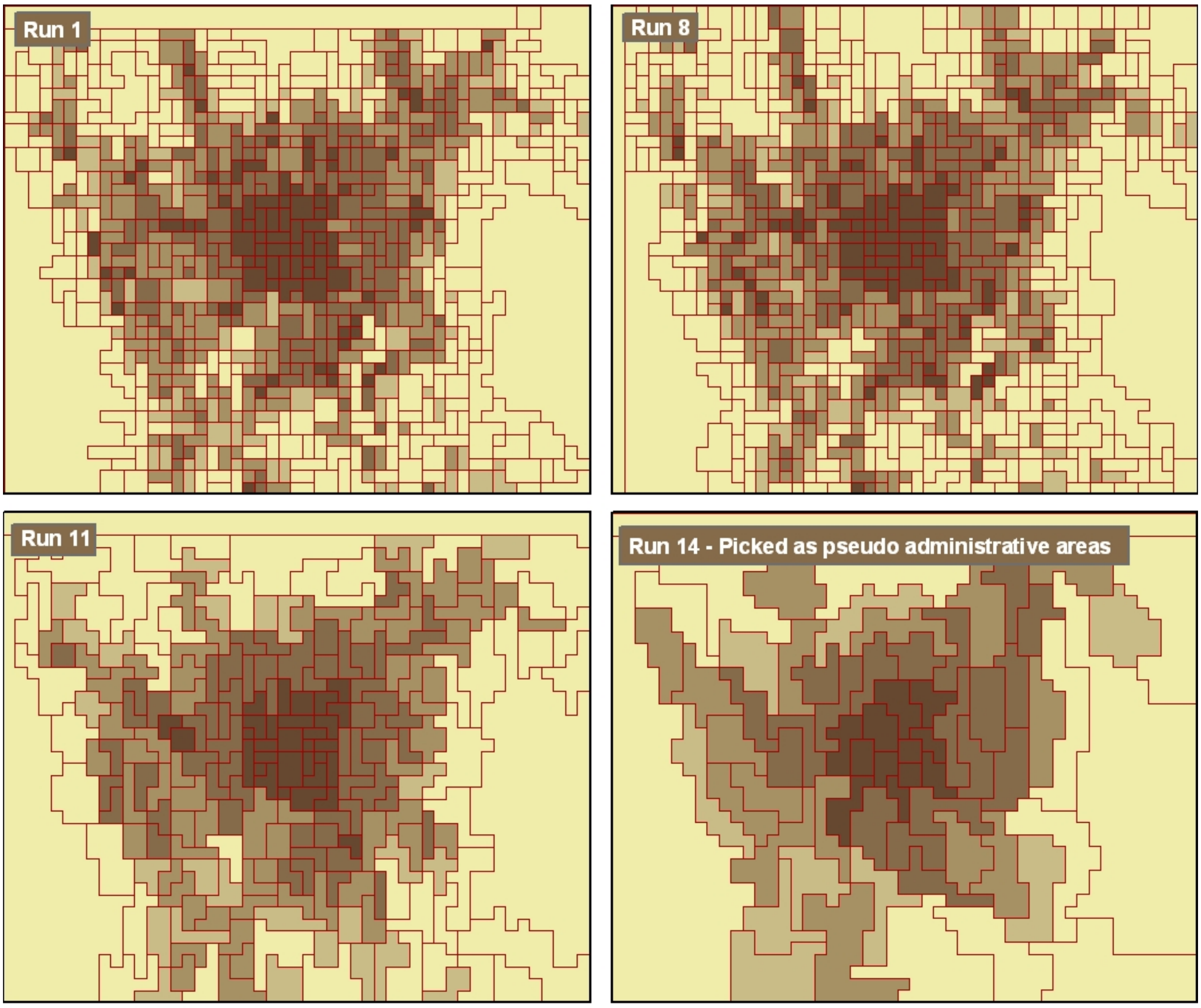
Most spatial epidemiological studies use administrative boundaries as units of analysis due to the availability of census and socio-economic data at these levels. Wards are a mid-level administrative unit frequently used because they are small enough to reflect local variations in exposures and disease rates and achieve a relatively homogeneous study population (*Briggs et al., 2007*). At the same time wards are large enough to contain enough people to provide the necessary statistical power for epidemiological studies and to avoid the so-called 'small number problem' which introduces uncertainty in the estimation of risks (*Elliott et al., 2007; Lawlor et al., 2005*). The creation of pseudo administrative areas for SIENA is therefore based on average population and area values of wards.

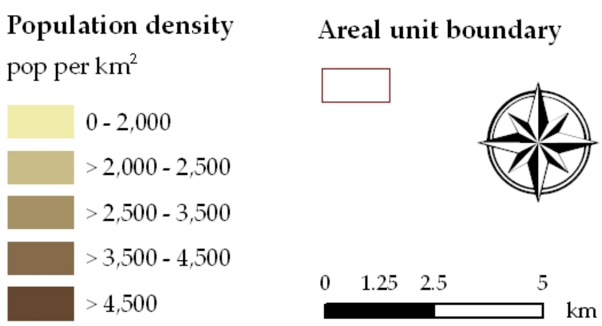
The Automated Zone Matching program (AZM) is used to define these pseudo administrative areas (*Martin, 2003*). The AZM algorithm was originally designed to create areal units for zone matching by achieving an optimal match between two input geographies. Being based on Openshaw's Automated Zoning Procedure (AZP) algorithm, however, it can also be used to create new zone systems (*Openshaw, 1977*). AZP iteratively recombines building blocks into output areas aiming to maximise a value of an objective function.

The AZM program requires two input files containing topographical details of the input geographies. The first file holds information on the intersection of the building blocks' polygons and the second file information on their contiguity. The two input files are created following the instruction from the AZM help file. Ideally, the 25 x 25 m grid cells of SIENA should be used as building blocks for the pseudo administrative areas but AZM has an input limit of 2000 zones. In order to fulfil this requirement, the 25 x 25 m grid cells are aggregated to 275 x 275 m grid cells. These 275 x 275 m grid cells form the building blocks of the pseudo administrative areas.

The creation of new zones within the AZM program is constrained by zoning rules. These rules give the user control over the parameters used by the AZM program in creating the new zones. They include information on the population threshold, i.e. the minimum population in the output zones, on the population target, i.e. the ideal population size for the output zones, as well as a shape control, which defines the compactness of an output area.

AZM will run a user specified number of iterations. A random input area is chosen at each iteration, it is swapped into an output area and the effect on the overall performance given the zoning controls considered. Swaps, which improve the overall requirements, are retained. The best solution after the specified number of iterations is presented as the new output zone. The program is run with Simulated Annealing (SA) which allows the program to accept swaps during the first half of the run that can reduce the overall reduction in the suitability of the simulation in order to improve overall performance of the final solution.





Run	Population density (pop/km²) min	mean	max	stdev
1	13	3,003	16,450	2,168
8	16	2,935	16,450	2,935
11	76	2,716	15,617	3,605
14	244	3,760	13,088	2,763
CASWards (urban)	21	5,618	70,046	2,745

Figure 16. Four examples of different pseudo administrative boundaries created with AZM visualising population densities

As mentioned above, ward specifications are used to inform the AZM model. Census area statistic wards (CASWards) in urban areas contain a minimum number of 100 people and have on average 7,400 people living within their boundaries (population max: 35,102; Stdev: 4,164). Their size in urban areas varies between 0.07 km^2 and 153.70 km^2 (area mean: 5.12 km^2; Stdev: 7.85 km^2). The pseudo administrative areas created with the AZM program are designed to reflect these CASWard specifications. Various AZM runs are performed using different zoning rules in order to approach this target (Table 4).

Table 4. Design scenario specifications and scenario statistics for the different runs performed with the AZM

Run	Population threshold	Population target	Shape control	Weight target	Population mean	Population stdev	Area mean	Area stdev
1	100	7,400	1	1	415	396	0.20	0.70
2	100	7,400	1	5	409	378	0.20	0.67
3	100	7,400	1	10	414	492	0.20	0.69
4	100	7,400	0.5	10	412	464	0.20	0.68
5	100	7,400	2	10	411	446	0.20	0.68
6	100	7,400	5	10	406	365	0.20	0.67
7	100	7,400	10	10	411	391	0.02	0.70
8	100	7,400	25	25	408	296	0.20	0.58
9	100	7,400	0.5	1.5	413	394	0.20	0.66
10	100	15,000	2	10	421	400	0.21	0.68
11	1,000	15,000	2	20	1,345	634	0.66	1.48
12	1,000	20,000	25	50	1,339	542	0.66	1.46
13	2,500	25,000	10	50	3,208	775	1.59	2.57
14*	3,500	30,000	10	50	4,529	1,071	2.23	3.11
15	3,500	30,000	10	100	4,666	1,096	2.30	3.51

* Boundaries from run 14 are chosen as pseudo administrative areas for SIENA

All design scenarios are run with 100 iterations in order to allow enough scope to improve the output zones. The results of the AZM runs are converted into ESRI coverages and then explored visually and statistically. The suitability of the output geography as pseudo administrative areas for SIENA is assessed primarily on the population numbers in each areal unit and realistic shapes, important factors when choosing units of analysis in spatial epidemiology (see Figure 16). The output zones of the run with the most compact shapes and specifications closest to the CASWard requirements (run 14) are chosen as the pseudo administrative area for the urban simulation.

Indoor air pollution

Awareness of indoor air pollution as a public health concern is increasing as more studies provide evidence that health effects from indoor air pollution can be as high as those from outdoor air pollution (*Bernstein et al., 2008; Samet, 1993; Spengler & Sexton, 1983)*. Time-activity studies have shown that people tend to spend most of their time indoors and are therefore mostly exposed to indoor air (*Nelson et al., 1994*). Obtaining information about indoor concentration levels, thus, becomes more and more the focus of exposure assessments especially as indicators for personal exposure. Indoor air pollution arises from various sources both indoor and outdoor. Apart from obtaining direct measurements it is difficult to establish and few attempts have been made to model indoor air pollution for epidemiological studies.

In order to estimate indoor air pollution levels for the urban simulation a newly developed stochastic mass-balance model (INDEX) is used (*Jamieson, 2010*). The INDEX model computes indoor concentration levels by simulating the penetration of outdoor air into a building or room and the behaviour of pollutants once indoors. Both processes are largely influenced by the ventilation and deposition characteristics of the building, a fact that is highlighted in the INDEX model.

In order to estimates indoor concentrations using the INDEX model, outdoor concentrations, penetration factors, deposition rates and air exchange rates have to be established for each indoor location. Data from SIENA is used to populate the INDEX model; in case of missing data the default options of the model are used. Table 5 lists the main INDEX parameters and their source. Indoor concentrations are calculated for each residential address using the modelled hourly air pollution concentrations as input.

Table 5. INDEX model parameters

Parameter	Description	Source data & computation
Outdoor concentration of pollutant *s* (*Cs*)	Measured concentration of pollutant *s* at time *t*	Hourly concentration of pollutant *s* at indoor location *x*, derived data from SIENA
Penetration factor (*P*)	Proportion of incoming pollutant *s* that is filtered during ingress	Default option: 0.05
Deposition rate (β)	Number of times per hour that pollutant *s* would be totally lost via deposition from air via deposition	$\beta = \frac{G_s \times \left((0.1086 \times S_s^2) + (0.0312 \times S_s) + 0.0052\right)}{Rh - Mh}$ where G_s is the specific density of pollutant *s* computed for location *x* as a random number between 0.8 (soot) and 3.5 (mineral particles) (*Coudray et al., 2008),* S_s S_s is the median diameter of individual particles computed for location *x* as a random number between 1.5 and 2.5 μm (*Coudray et al., 2008*), *Rh* is the room height as specified in the building characteristics of the urban simulation for the location *x*, *Mh* is the monitoring height (default option: 1 metres).

Air exchange rate (λ)	Frequency of complete air exchange via wall *i* in time *t*	$\lambda = 3600 \times T \times \ (0.0025 \times Kd \times Wa_i \times Uf_{ti}) \times \frac{Ptot_{it}}{Rv}$ where T is the time step for each calculation set to 1 hour, Kd is the wind direction coefficient computed as: $Kd = 0.75 + \left[0.25 \times \cos\left(Zf_{ti} \times \frac{\pi}{180}\right)\right]$ Where Zf_{ti} is the relative wind direction relative to orientation of wall *i* as specified in the building characteristics of the urban simulation for location x, Wa_i is the area of external wall i computed as: $Wa_i = Rh \ \times \frac{Ra}{2}$ where *Ra* is the room area and *Rh* the room height as specified in the building characteristics of the urban simulation for location x, Uf_{ti} is the estimated wind speed at wall *i* at time *t* computed as: $Uf_{ti} = Um_t \times Kr \times Bh^{Kv}$ where Um_t is the measured wind speed at time *t*, defined as the average wind speed in the urban simulation, *Kr* is the index of surface roughness (default option for urban areas: 0.21), *Bh* is the building height as specified in the building characteristics of the urban simulation for location x and *Kv* is the vertical wind speed coefficient (default option of urban areas: 0.33) , $Ptot_{it}$ is the total porosity of wall *i* at time *t* computed as: $Ptot_{it} = Pn_i + Pa_{it}$ where Pn_i is the natural porosity of wall *i* computed as random number between 0.1 (default option) and Pa_{it}; Pa_{it} is the porosity via controlled inlets through wall *i* at time *t* computed as $\frac{Ia_{it}}{Wa_i}$, which are specified in the building characteristics of the urban simulation for location x, Rv is the room volume computed as: $Rv = Ra \times Rh$ where *Ra* is the room area and *Rh* the room height as specified in the building characteristics of the urban simulation for location x.

The examples of modelling derived data described above show how SIENA can be used generate new data using both established models such as ADMS-Urban or the AZM program as well as newly developed models such as the INDEX model. Comparisons of the modelled pollutant concentrations in SIENA with real-world monitoring data, for example, show that using the existing data structure of SIENA to run established models will result in expedient results. On the other hand, the example of the INDEX model to simulate indoor air pollution concentrations shows how, given applicable contextual data, newly developed models can be run and tested. The INDEX model, so far, has only been a theoretical model without practical application due to the data requirements. But SIENA provides a platform to simulate the necessary input data and, therefore, to apply the INDEX for the first time. This will give the authors of the model a chance to evaluate its performance and assess its behaviour in a practical situation. Equally, this raises questions as to the accuracy of the model output obtained from the INDEX models and results of any application using these outputs have to be treated with the necessary wariness.

4. Discussion

4.1 SIENA considerations

The purpose of SIENA is to provide a user-controlled system in which to simulate and explore the ways in which environmental factors affect human exposure to risk factors and how these might change under different scenarios. This objective is the rationale behind the development process of SIENA described in this chapter. Each decision along the process is driven by this purpose, be it decisions on the data to be included, decisions on how this data is modelled and incorporated or decisions on the overall structure of SIENA.

The development of SIENA is based on three main conceptual conditions as described in the introduction, all of which reflect the above mentioned objective: firstly, SIENA should represent a medium-sized city in GB and reflect especially the real-world urban structure; secondly, SIENA should preserve the observed real-world complexities with special focus on the interactions between the different urban components; and thirdly, SIENA should allow a high degree of flexibility to enable the implementation of different scenarios and consequently the development of scenario specific data whilst observing the two previously mentioned conceptual conditions.

The first conceptual condition, the representation of a medium-sized city is achieved by the pooling of structural patterns and interactions observed in the sample cities. The statistical analysis of the sample cities identified the merging of the sample cities' characteristics as the best approach to implement a representative urban area. Extracting similar elements and patterns of the urban components in the sample cities and applying these common features to SIENA allows concentrating on features common to all sample cities and to ignore local discrepancies. The quality assessment carried out supports this assumption. Averaging and combining the spatial information and characteristics observed in the sample cities does result in realistic structural patterns and urban features in SIENA.

The second conceptual condition, the preservation of urban complexity, is realised through the definition of design rules to build SIENA. The observing of these design rules allows focusing on the quantity and quality of structural elements as well the preserving of as much real-world complexity of spatial interactions as feasible.

The principle of defining averaging rules based on sample cities rather than determining fixed structures accompanies the third conceptual condition, the retaining of flexibility. Flexibility is also the rationale behind the three-tier data structure of SIENA. Core data, contextual data and derived data from three distinct data types, each with its own rational, purpose and function. Core data build the foundation for the other two data types and great care was taken to represent topography, the transportation network, land cover, and population distribution as realistically as possible. The quality assessment carried out comparing SIENA characteristics to the sample cities indicates that this has been achieved. Both, the physical features as wells the spatial interactions between the core data show similar patterns in SIENA and in the sample cities. If implementing new scenarios in SIENA, scenario-specific contextual data can be added as it is or based on the structural patterns of the core data. Scenario-specific contextual data developed based on core data should ensure the maintenance of the urban complexities. Again, the validations carried out support this assumption. SIENA can also be used as a data platform to derive newly simulated data based on core and contextual data. Results of the derived data generated using both core and contextual data as input are another indication that the SIENA urban structure behaves in realistic ways. Given that the model input is based on contextual data modelled in compliance with the SIENA requirements, SIENA produces realistic model outputs.

When introducing contextual data to SIENA without assessing real-world behaviour first, spatial information and accuracy loss as well as the introduction of additional uncertainty is inevitably. All decisions taken to further develop SIENA or apply it in a scenario context have to be backed by observations of pattern and functions in the real-world. One exception is the incorporation of contextual data without any spatial considerations as is demonstrated with the meteorological data. Putting SIENA in a generally wider context, as is the case with, for example, prevailing wind direction in an urban area does not need to be supported by intra-urban structural analysis. But intra-urban variability would have to be analysed if, for example, introducing wind direction measured at various monitoring stations within SIENA.

As already mentioned, the construction process for SIENA is defined by the objective to allow enough flexibility for the user to adapt SIENA to various scenarios but still be as realistic as possible in modelling the urban complexities. This objective is achieved by building SIENA at a very fine 25 x 25 m resolution. This resolution is on the one hand detailed enough to detect variation for small area analysis and analysis at household level (25 x 25 m gridded points represent residential addresses) and on the other hand flexible enough to aggregate to lower resolution if required by a scenario.

As has been shown, the approach taken to develop SIENA is strongly function driven rather than design driven. This principle is further underlined by the choice of model to generate the different data types. No assumptions or assessments are made about the models used. In each case, the implemented models are the ones that are most suitable for the given nature of the data, to achieve the scenario-specific data requirements in term of data detail and resolution, to comply with the design rules and to allow as much flexibility as possible for the user.

The detail required, for example, in case of minor road density results in a different approach than for the main roads. For future scenarios, the minor road density could be increased or even, if the necessity arises, minor roads could be modelled as line features using, for example, the construction approach. The probabilistic approach adopted to model some of the core data and to a lesser degree the contextual data is one of the models used in SIENA that allows the most flexibility. As soon as new urban data structures are introduced, previous modelled data sets can be remodelled including the new features to recalculate the allocation probabilities. As has been shown, the probabilistic model can also be adapted to suit the needs and specifics of a variety of different data. One example is adopting a hierarchical approach when modelling the population density, first distributing it to land cover patch level using weights and then using the probabilistic model to calculated population density at the 25 x 25 m grid level. SIENA, therefore, provides the flexibility to adapt models according to the data needs.

Different model approaches are used for different data types. Restricting the model to one approach, for example, the cellular automata approach or the probabilistic model approach would result in a too confined paradigm which would not allow flexibility in coherence with the conceptual conditions. The scope of applied models consists of techniques easily implemented in widely-used programs such as a statistical packages and GIS software in order to allow the SIENA user to easily follow the modelling techniques introduced here if simulating scenario-specific contextual data. This would not be the case if very complex, data and time intensive and specialised models would have been used to model SIENA.

SIENA simulates an urban area and general tendencies or trends can be provided using SIENA. But no absolute results will be obtained when applying SIENA for analysis and scenarios. The fact that SIENA is based on an amalgamation of cities helps to ensure that its results have general applicability. Nevertheless it needs to be remembered that SIENA is a model and, as with all models, is therefore not directly representative of every real-world setting. Instead results need to be interpreted as indicative of what is likely to happen in the real-world.

All simulations of course bear the problem of validation and how to assess uncertainty. SIENA is only a representation of reality, a model, and therefore a loss of complexity and introduction of uncertainty is inevitable. No satisfying procedure exists to thoroughly validate the simulation and assess uncertainty. One way chosen here is the quality assessment which, as mentioned before, shows that the SIENA urban structure and interactions are comparable to the sample cities. Another way is via behavioural analysis. Does SIENA behave in the same way as real-world urban areas would? This can only be properly assessed if more and more applications are carried out with SIENA and the observed behaviours compared to the real-world.

One example, hereof, is the application of a land use regression model (LUR), a model with a long history in environmental epidemiology which has been used in many urban areas in the past (*Briggs et al., 2000; Hoek et al., 2008b; Wheeler et al., 2008)*. In LUR, the surroundings of a small number of monitoring sites are described by predictor variables such as land use, road network, meteorological data and terrain. Linear regression functions are applied in order to establish a relationship between monitored concentrations and predictor variables and the model is then applied to unsampled locations in the study area (*Hoek et al., 2008a)*. Comparing results of the LUR model implemented in SIENA to that of LUR models conducted in the real-world gives an indication of how well the SIENA data structure behaves. Applying the LUR to the SIENA data results in a model R^2 of 0.64 between modelled and observed concentrations for validation sites. A study estimating exposure to traffic related air pollution in Sheffield, one of the sample cities, achieved an average LUR model R^2 of 0.68 (*de Hoogh, 1999*). These comparable model results present only an indication that the SIENA data structure behaves in similar ways than real-world cities and is no definite validation.

SIENA is built on an application-driven approach. It will be extended, expanded, more complexity added and built-on with each scenario, case study or application carried out. Confidence will grow as more and more applications demonstrate that observed behaviours are similar or comparable to real-world phenomena. This will be the true validation of the model, the application of more and more scenarios.

Reference List

ABRAHAM J.M. & HENDERSHOTT P.H. (1996) Bubbles in metropolitan housing markets. *Journal of Housing Research* **7**, 191-207.

ALAVANJA M.C.R., WARD M.H. & REYNOLDS P. (2007) Carcinogenicity of agricultural pesticides in adults and children. *Journal of Agromedicine* **12**, 39-56.

ALPEROVICH T. & SOPASAKIS A. (2008) Stochastic Description of Traffic Flow. *Journal of Statistical Physics* **133**, 1083-1105.

BALK D., MCGRANAHAN G. & ANDERSON B. (2008) Urbanization and ecosystems: Current patterns and future implications. In *The new global frontier: Urbanization, poverty and environment in the 21st century*. Eds G.Martine, G.McGranahan, M.Montgomery & R.Fernandez-Castilla. Earthscan. pp. 183-201.

BERNSTEIN J.A., ALEXIS N., BACCHUS H., BERNSTEIN I.L., FRITZ P., HORNER E., LI N., MASON S., NEL A., OULLETTE J., REIJULA K., REPONEN T., SELTZER J., SMITH A. & TARLO S.M. (2008) The health effects of nonindustrial indoor air pollution. *Journal of Allergy and Clinical Immunology* **121**, 585-591.

BRIGGS D., DE HOOGH C. & GULLIVER J. (2010) Comparative assessment of GIS-based methods and metrics for modeling exposure to air pollution. *Journal of Toxicology and Environmental Health* **in press**,

BRIGGS D.J., DE HOOGH C., GULLIVER J., WILLS J., ELLIOTT P., KINGHAM S. & SMALLBONE K. (2000) A regression-based method for mapping traffic-related air pollution: application and testing in four contrasting urban environments. *Science of the Total Environment* **253**, 151-167.

BRIGGS D.J., FECHT D. & DE HOOGH C. (2007) Census data issues for epidemiology and health risk assessment: experiences from the Small Area Health Statistics Unit. *Journal of the Royal Statistical Society Series A-Statistics in Society* **170**, 355-378.

BRUNEKREEF B. & HOLGATE S.T. (2002) Air pollution and health. *Lancet* **360**, 1233-1242.

BRUNT D. (2007) *Meteorology*. Karig Press.

CARR D., VON EHRENSTEIN O., WEILAND S., WAGNER C., WELLIE O., NICOLAI T. & VON MUTIUS E. (2002) Modeling annual benzene, toluene, NO2, and soot concentrations on the basis of road traffic characteristics. *Environmental Research* **90**, 111-118.

CERC (2006) *ADMS-Urban: An urban air quality management system - User guide*. Version 2.2 Ed. Cambridge:

CERC (2008) *EMIT: Atmospheric Emissions Inventory Toolkit - User guide*. Version 2.3 Ed.

CONGALTON R.G. (1991) A review of assessing the accuracy of classifications of remotely sensed data. *Remote Sensing of Environment* **37**, 35-46.

COUDRAY N., DIETERLEN A., VIDAL L., ROTH E., TROUVE G. & BISTAC S. (2008) Image processing nanoparticle size measurement for determination of density values to correct the ELPI measures. *Precision Engineering-Journal of the International Societies for Precision Engineering and Nanotechnology* **32**, 88-99.

DE HARTOG J.J., AYRES J.G., KARAKATSANI A., ANALITIS A., BRINK H., HAMERI K., HARRISON R., KATSOUYANNI K., KOTRONAROU A., KAVOURAS I., MEDDINGS C., PEKKANEN J. & HOEK G. (2010) Lung function and indicators of exposure to indoor and outdoor particulate matter among asthma and COPD patients. *Occupational and Environmental Medicine* **67**, 2-10.

DE HOOGH C. (1999) Estimating exposure to traffic-related pollution within a GIS environment. *PhD Thesis, University of Leicester*

DE MARCO R., MARCON A., RAVA M., CAZZOLETTI L., PIRONI V., SILOCCHI C. & RICCI P. (2010) Proximity to chipboard industries increases the risk of respiratory and irritation symptoms in children The Viadana study. *Science of the Total Environment* **408**, 511-517.

DEPARTMENT FOR TRANSPORT (2006) *Transport statistics bulletin - Road statistics 2006: traffic, speeds and congestion*.

DIMITROULOPOULOU C., ASHMORE M.R., HILL M.T.R., BYRNE M.A. & KINNERSLEY R. (2006) INDAIR: A probabilistic model of indoor air pollution in UK homes. *Atmospheric Environment* **40**, 6362-6379.

DOCKERY D.W., POPE C.A., III, XU X., SPENGLER J.D., WARE J.H., FAY M.E., FERRIS B.G., Jr. & SPEIZER F.E. (1993) An association between air pollution and mortality in six US cities. *New England Journal of Medicine* **329**, 1753-1759.

ELLIOTT P., BRIGGS D.J., MORRIS S., DE HOOGH C., HURT C., JENSEN T.K., MAITLAND I., RICHARDSON S., WAKEFIELD J.C. & JARUP L. (2001) Risk of adverse birth outcomes in populations living near landfill sites. *British Medical Journal* **323**, 363-368.

ELLIOTT P., SHADDICK G., WAKEFIELD J.C., DE HOOGH C. & BRIGGS D.J. (2007) Long-term associations of outdoor air pollution with mortality in Great Britain. *Occupational and Environmental Medicine*

ESRI (2001) *ArcGIS Spatial Analyst: Advanced GIS spatial analysis using raster and vector data*. Redlands: ESRI.

GAUDERMAN W.J., AVOL E., LURMANN F., KUENZLI N., GILLILAND F., PETERS J. & MCCONNELL R. (2005) Childhood asthma and exposure to traffic and nitrogen dioxide. *Epidemiology* **16**, 737-743.

GHAZALI N.A., RAMLI N.A., YAHAYA A.S., YUSOF N.F.F.M., SANSUDDIN N. & AL MADHOUN W.A. (2010) Transformation of nitrogen dioxide into ozone and prediction of ozone concentrations using multiple linear regression techniques. *Environmental Monitoring and Assessment* **165**, 475-489.

GOYAL P. & MITTAL R. (2004) Line source models to estimate urban air concentration in different wind conditions. *Traffic and Transportation Studies, Proceedings* 602-613.

HOEK G., BEELEN R., DE HOOGH C., VIENNEAU D., GULLIVER J., FISCHER P. & BRIGGS D. (2008a) A review of land-use regression models to assess spatial variation of outdoor air pollution. *Atmospheric Environment* **42**, 7561-7578.

HOEK G., BEELEN R., PEBESMA E., VIENNEAU D., DE HOOGH C. & BRIGGS D. (2008b) Mapping of air pollution at a fine spatial scale across the European Union. *Science of the Total Environment* **407**, 1852-1867.

JAMIESON S. (2010) Spatio-temporal modelling of indoor particulate matter in naturally-ventilated buildings. *PhD thesis, Imperial College London*

JENNESS J. & WYNNE J.J. (2005) Kappa analysis (kappa_stats-avx) extension for ArcView 3.x. *Jenness Enterprises*

JENSEN J.R. (1996) *Introductory digital image processing: a remote sensing perspective*. 2nd Ed. Upper Saddle River, New Jersey, USA: Prentice Hall, Inc.

KRAJCOVICOVA J. & ESCHENROEDER A.Q. (2007) Comparative health risks of domestic waste combustion in urban and rural Slovakia. *Environmental Science & Technology* **41**, 6847-6853.

KUTTLER W. (2008) The urban climate: basic and applied aspects. In *Urban ecology: An international perspective on the interaction between humans and nature*. Eds J.M.Marzluff, E.Shulenberger, W.Endlicher, M.Alberti, G.Bradley, C.Ryan, U.Simon & C.ZumBrunnen. Springer. pp. 233-248.

LAWLOR D.A., SMITH G.D., PATEL R. & EBRAHIM S. (2005) Life-course socioeconomic position, area deprivation, and coronary heart disease: Findings from the British women's heart and health study. *American Journal of Public Health* **95**, 91-97.

LONGLEY P. & BATTY M. (1996) *Spatial analysis: modelling in a GIS environment*. John Wiley & Sons, Inc.

MARTIN D. (2003) Extending the automated zoning procedure to reconcile incompatible zoning systems. *International Journal of Geographical Information Science* **17**, 181-196.

MCHUGH C.A., CARRUTHERS D.J. & EDMUNDS H.A. (1997) ADMS-Urban: an air quality management system for traffic, domestic and industrial pollution. *International Journal of Environment and Pollution* **8**, 666-674.

MITCHELL C.S., ZHANG J.F.J., SIGSGAARD T., JANTUNEN M., LIOY P.J., SAMSON R. & KAROL M.H. (2007) Current state of the science: Health effects and indoor environmental quality. *Environmental Health Perspectives* **115**, 958-964.

NAZAROFF W.W. (2004) Indoor particle dynamics. *Indoor Air* **14**, 175-183.

NELSON W.C., OTT W.R. & ROBINSON J.P. (1994) The national human activity pattern survey (NHAPS): use of nationwide activity data for human exposure assessment. *Paper number 94- WA75A.01 presented at the A&WMA 87th Annual Meeting: Cincinnati, OH, 1994 and EPA Report, 600/A-94/147, AREAL.Research Triangle Park, NC*

NETCEN (2009) *Air pollution in the UK: 2008*.

NICO G., RUTIGLIANO P., BENEDETTO C. & VESPE R. (2005) Terrain modelling by kinematical GPS survey. *Natural Hazards and Earth System Sciences* **5**, 293-299.

OPENSHAW S. (1977) Geographical solution to scale and aggregation problems in region-building, partitioning and spatial modeling. *Transactions of the Institute of British Geographers* **2**, 459-472.

PARKER D.C., MANSON S.M., JANSSEN M.A., HOFFMANN M.J. & DEADMAN P. (2003) Multi-agent systems for the simulation of land-use and land-cover change: A review. *Annals of the Association of American Geographers* **93**, 314-337.

ROSENLUND M., FORASTIERE F., STAFOGGIA M., PORTA D., PERUCCI M., RANZI A., NUSSIO F. & PERUCCI C.A. (2008) Comparison of regression models with land-use and emissions data to predict the spatial distribution of traffic-related air pollution in Rome. *Journal of Exposure Science and Environmental Epidemiology* **18**, 192-199.

SAMET J.M. (1993) Indoor air pollution: a public health perspective. *Indoor Air* **3**, 219-226.

SANTE I., GARCIA A.M., MIRANDA D. & CRECENTE R. (2010) Cellular automata models for the simulation of real-world urban processes: A review and analysis. *Landscape and Urban Planning* **96**, 108-122.

SPENGLER J.D. & SEXTON K. (1983) Indoor Air-Pollution - A Public-Health Perspective. *Science* **221**, 9-17.

VAN AALST R. (2003) *The contribution of traffic to concentrations of PM10 in streets and urban areas*. European Environment Agency.

WALLACE L. (1996) Indoor particles: A review. *Journal of the Air & Waste Management Association* **46**, 98-126.

WANG F., KETZEL M., ELLERMANN T., WAHLIN P., JENSEN S.S., FANG D. & MASSLING A. (2010) Particle number, particle mass and NOx emission factors at a highway and an urban street in Copenhagen. *Atmospheric Chemistry and Physics* **10**, 2745-2764.

WHEELER A.J., SMITH-DOIRON M., XU X., GILBERT N.L. & BROOK J.R. (2008) Intra-urban variability of air pollution in Windsor, Ontario - Measurement and modeling for human exposure assessment. *Environmental Research* **106**, 7-16.

ZAMBON P., RICCI P., BOVO E., CASULA A., GATTOLIN M., FIORE A.R., CHIOSI F. & GUZZINATI S. (2007) Sarcoma risk and dioxin emissions from incinerators and industrial plants: a population-based case-control study (Italy). *Environmental Health* **6**,

ZUURBIER M., HOEK G., OLDENWENING M., LENTERS V., MELIEFSTE K., VAN DEN HAZE P. & BRUNEKREEF B. (2010) Commuters' Exposure to

Particulate Matter Air Pollution Is Affected by Mode of Transport, Fuel Type, and Route. *Environmental Health Perspectives* **118**, 783-789.