SOME IDEAS FOR INTEGRATING MULTIDISCIPLINARY SPATIAL DATA

W. Shi *, L. Meng

Technical University of Munich, Department of Cartography, 80333 Munich, Germany –
(wei.shi, liqiu.meng)@bv.tum.de

Commission VI, WG VI/4

KEY WORDS: Data integration, data mining, data matching, gridding, mathematic morphology, physical models

ABSTRACT:
Unlike the integration of geospatial data that deals with different geometric expressions, resolutions and actualities etc., integrating different domain datasets associated with the same geo-reference must consider additionally the specific domain knowledge and models. A number of spatial reference datasets and multidisciplinary datasets are selected as test examples. An analysis of data quality is conducted with regard to the resolution, acquisition method, format and reliability etc. On the basis of data matching, gridding, mathematic morphology and physical modelling, the authors introduce different concepts of data integration with the purpose to map non-geometric semantic attributes onto a uniform geo-space. Such kind of complicated integration is a necessary preparation for meaningful knowledge discovery and data mining.

1. INTRODUTION

1.1 Background

With the start of the first German civilian radar satellite system TerraSAR-X in the summer 2006, remote sensing data with a high resolution of up to one meter and a daily actuality could achieve an optimised quality of satellite products and open up a large number of new applications including the one reported in this paper.
To bring the TerraSAR-X data into use, three processing steps are necessary: (1) converting radar signals into raster images, (2) extracting geo-objects from raster images, possibly in combination with other sensory data, and coding the results in vector formats, (3) discovering useful and interesting knowledge from vector objects and coding the result in form of qualitative statements, quantitative measures or association rules. One of the essential techniques along this pipeline is data mining of different levels.
Object extraction from sensory data, i.e. the base level of data mining, aims at detecting the identities of individual geo-objects in the real world in terms of their spatial and a small part of non-spatial attribute values. For example, from precise and actual radar data, it is possible to detect vehicles along a road from which microscopic information such as traffic volume, noise and waste gas emission for a given road and moment can be derived and compared with the macroscopic information, i.e. statistically observed values along the same road. The outcome of current extraction processes from remote sensing data remains mainly spatial attributes from which locations, shapes, sizes, topological relations and distributional characteristics of objects can be derived and analysed. However, a database composed of purely spatial aspects, no matter how large it could be, hardly allows the meaningful data mining of higher levels. For example, the rule "if a Canadian town is large and adjacent to a large water body, then it is close to the U.S. border with the possibility of 78%" is of little value [Han, 1997]. On the other hand, a data mining procedure conducted on purely non-spatial attributes of geo-objects may lead to the discovery of interesting rules describing the semantic correlations among these objects. However, the equally interesting interactions between spatial and non-spatial aspects remain untouched [Wang 1997, Sheikholeslami 2000]. For example, a number of population migration rules can be discovered on the basis of demographic attribute values. Yet, more essential migration rules might be still hidden because the population movement coheres strongly with the traffic conditions, living situation, job market, urban security etc.
Data mining has experienced a fast development that reached its peak the 90's. Most research work has emphasized on the efficiency and robustness of algorithms. However, how many useful patterns a certain algorithm can discover largely depends on the database itself. Under the general assumption that the more data items and attributes a database contains, the more likely there would be interesting interrelationships, this paper is devoted to the issue of data integration, aiming at bringing spatial and non-spatial attributes of geo-objects together.

1.2 Data integration

The successful stories of data integration reported so far are mostly concentrated on uniting geo-objects by means of their spatial attributes that are resulted from different acquisition methods and stored in different formats such as raster imagery, vector maps etc. [Yang 1992], [Shekhar 2003]. In spite of their heterogeneity in terms of resolution, scale range, actuality, reliability, completeness, consistency and so on, the geo-objects to be compared do share much spatial commonality with regard to their locations, general shapes, extents and surface structures. Therefore, a good or very good matching is possible within certain tolerance thresholds in a given context. The integrated spatial data usually have an improved actuality, reliability, completeness and consistency. However, the integration of non-spatial attributes so far has been mostly conducted on aggregated statistical data at a too low semantic resolution for the subsequent data mining work.

* Corresponding author.

Being stimulated by the accessibility of TerraSAR-X data at high resolution and actuality, this paper attempts to geo-reference and integrate non-spatial multidisciplinary data at the address level.

2. DATA SOURCES

To test our integration ideas, the following data sets related to Greater Munich, i.e. city Munich and its suburb have been collected:

(1) Basic Digital Landscape Model (Basis-DLM) from German mapping agencies
(2) Street data from TeleAtlas Corp.
(3) Postal addresses from German Federal Post Office
(4) Traffic load database from the Traffic Office Munich
(5) Traffic noise from Bavarian Ministry for Environment and Healthiness
(6) Standard ground value (raster) from Advisory Committee for land value of the city Munich
(7) Demographic data from Statistic Agency of city Munich
(8) Commuter data from Federal Employment Office

(1)–(3) serve as spatial references, whereas (4)–(8) are multidisciplinary thematic datasets to be integrated.

2.1 Spatial references

Our ideal primary spatial reference data - a seamless 3D cadastral model of city Munich overlaid on TerraSAR-X image - is not yet available. Therefore, the Basis-DLM in the scale range of 1:10,000 - 1:25,000 has been used instead for our work so far. It is created and updated through map digitization in combination with semiautomatic object extraction from imagery data. It contains geometric and semantic attributes describing topographic object types - settlement, road, hydrography, vegetation, boundaries, terrain, and a number of point features such as bus stops, cinemas, drugstores, etc. The data structure is defined in accordance with the terms of the Official Topographic Cartographic Information System (ATKIS). The values of the individual attributes are not completely available, for example, only a part of the street names are captured.

Our secondary spatial reference data - TeleAtlas street dataset - is created and updated through map digitization and GPS-supported field measurement. It contains precise road geometries, street names and navigation-relevant attributes including dynamic traffic information. The data structure is defined by TeleAtlas which is one of the most important data suppliers for car navigation systems.

The postal dataset is created and updated by postmen based on the road geometry of TeleAtlas dataset. The individual house numbers are stored as discrete points distributed along the two road sides. The location of each individual house number was estimated or interpolated by the postman based on the beginning and terminating house number delimiting each street segment and the known house-numbering rules. For this reason, it may deviate more or less from the true location of the corresponding house entrance, but the topological relationship is preserved.

2.2 Multidisciplinary thematic datasets

2.2.1 Traffic load and noise

The traffic load is represented as a matrix showing the ingress and egress points and the demand of traffic between them. Figure 1 shows the traffic load along the main streets of city Munich, where relative load measure is symbolized by the relative thickness of the road.

Figure 1. The traffic load of city Munich [source: Traffic Office Munich]

The traffic noise of city Munich as shown in Figure 2 is derived from the traffic load and mapped as different colours onto the corresponding streets. The influence of the road noise on inhabitants living nearby is not expressed. In this case, extending the linear distributing to the regional distribution is needed.

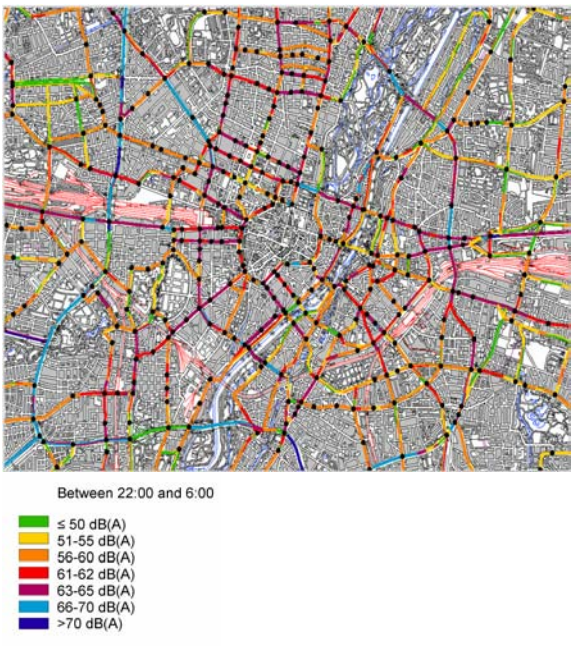


Figure 2. The road noise of the city Munich [source: Bavarian Ministry for Environment and Healthiness]

2.2.2 Standard ground value

Figure 3 illustrates the classified standard ground values in city Munich. A standard ground value or guiding land value is defined as an average land value per square meter for an area with essentially identical characteristics determined by purchase prices.

A standard ground value refers in principle to undeveloped land and bears no market values. Deviations of land in the worth-determining characteristics, which can be evaluated according to a variety of criteria such as kind and measure of the structural use, ground condition, the development condition and land organization cause deviations of its land value from the standard ground values.

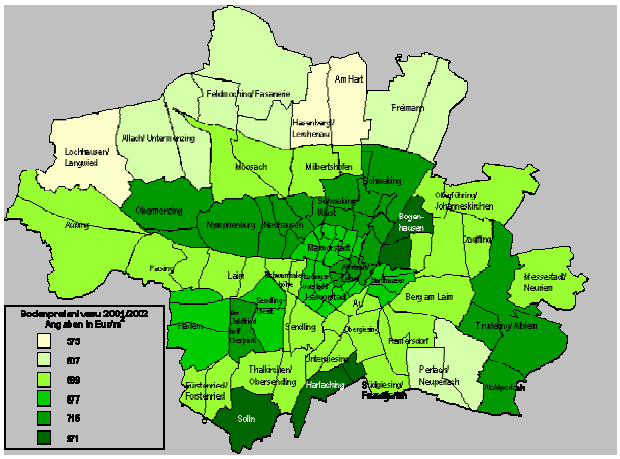


Figure 3. Standard ground values of the city Munich [source: Advisory Committee for land value of city Munich]

2.2.3 Demographic attributes and commuters

A part of demographic attributes is shown in Figure 4. In addition to the elementary properties of an inhabitant, address-related attributes of different granularity levels from urban district, sub-district, residential quarter, block of flats, street and house number are also stored in the dataset.

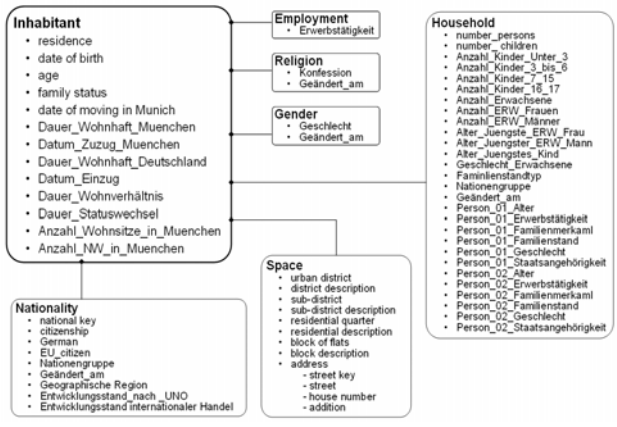
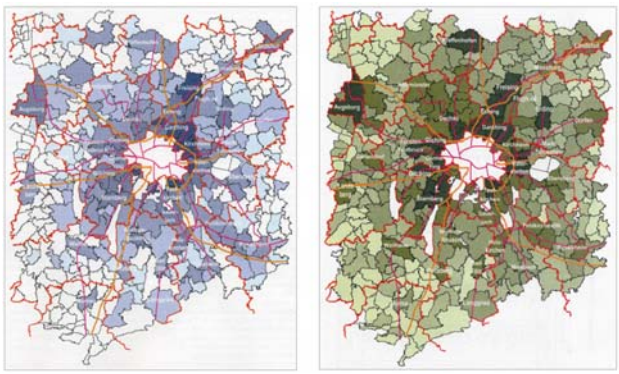


Figure 4. Demographic attributes [source: Statistic Agency of city Munich]

Commuter dataset is selected as it relates closely to traffic conditions, job market, living situation etc. Figure 5(a) presents the classified population of in-commuters who are living in suburb while working in city Munich, while Figure 5(b) indicates the out-commuters. In both maps, the darker the colour tone, the larger the corresponding number of commuters.

Normally, there are no exact numbers of the in- and out-commuters of a city. What are available are the sum and the geographical distribution of the commuters in terms of their social insurance. The total numbers of in- and out-commuters of Munich are currently estimated under the assumption: the ratio between employees liable to social insurance and those without social insurance for non-commuters is similar to that for commuters. Therefore, the numbers of employees liable to social insurance among in- and out-commuters can be proportionally extrapolated based on the officially estimated number of employees. The numbers of in- and out-commuters from and to each suburb commune can also be approximately extrapolated.



(a) (b)

Figure 5. The in- and out-commuters of the city Munich [source: Federal Employment Office]

3. CASE STUDIES OF DATA INTEGRATION

The objective of the integration is to fuse various datasets with technically different models onto a uniform metric space. However, there is no single integration strategy that can be applied everywhere. An insight into the complexity of the integration issue could be gained through the following case studies conducted in the Department of Cartography, Technical University of Munich.

3.1 Integration of postal addresses

Due to incomplete street names and missing addresses, the values of the current Basis-DLM is rather limited, especially with regard to the booming market of traffic- and transportation-related applications. It is therefore desirable to enrich the Basis-DLM with the available postal addresses from German Federal Post Office. However, the attempt of a direct matching failed because the post addresses are bundled with TeleAtlas road objects which reveal a different geometric and semantic resolution from that of Basis-DLM-dataset (see Figure 6 left). To get along with the problem, we divide the enrichment process into two main stages. The first stage is dedicated to the matching of road objects from Basis-DLM with their counterpart from TeleAtlas dataset. In the second stage, a projection based on the rubber sheet principle is established for each pair of matched road lines. Thus, any arbitrary point along a TeleAtlas road line can find its corresponding position along the homologous line in DLM-dataset. This enables a reasonable transfer of the discrete locations of house numbers from TeleAtlas to Basis-DLM (Figure 6 right). The technical details on the matching algorithm and matching results have been reported in [Zhang, et al. 2005].

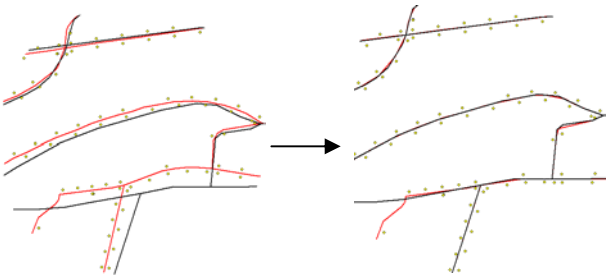


Figure 6. Overlapping of address points along streets in red from TeleAtlas with streets in black from Basis-DLM in black (left), address points are transferred onto streets of Basis-DLM after matching (right) [Zhang et al. 2005]

3.2 Integration of the standard ground value data

Many integration tasks need a conversion between vector and raster data which is in essence a conversion between discrete and continuous data. Applications such as spatial queries, thematic overlays, and statistical analysis require the direct access to discrete vector data that are object-wise stored in a database, while other applications may rely heavily on simpler and more flexible raster operations.

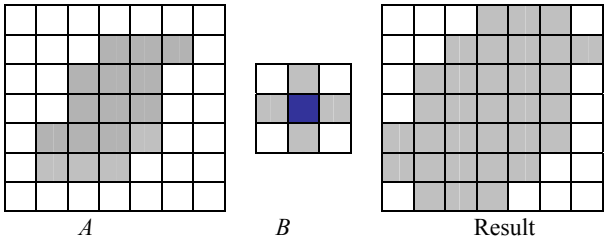
There are two basic raster data types: image data and grid data. A raster image is a two dimensional array of regularly spaced picture elements. The attribute information is embedded in colour tones grey values. The geometric and semantic resolution of a raster image is determined by its corresponding data recording sensors. The process of object exaction is usually conducted on a raster image and yields a vector database. Grid data is a more general type of raster data. It takes the form of a rectangular cell matrix aligned to the X and Y-axes overlaying an area. Each cell in the grid has the same size that corresponds to the resolution of the grid. This cell size can be determined artificially and on demand. In the simplest case, if the information about the bounding coordinates of the grid and the number of rows/columns is given, the location of each cell can be calculated. Therefore, no explicit location value is needed for each cell. Grid data typically stores attribute values for each cell in the grid. Standard ground value is one example for it.

In order to integrate grid-based standard ground value with vector-based objects in the Basis-DLM, we convert the latter into grid data at one-meter resolution by means of interpolation. There exist a great variety of interpolation algorithms such as those based on centre point, preponderant area, weight calculation, statistical surface, trend surface etc. [Goodchild 1980]. The subsequent integration is then a simple procedure of grid overlay. Though it is possible to convert the standard ground value into classes of vector areas, grid overlay proves more efficient for further computational operations aiming at data retrieval and knowledge discovery.

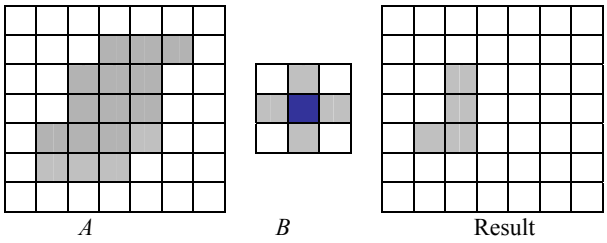
3.3 Integration of the emission data

3.3.1 Emission of exhaust gas

The emission from a road or a road section can be derived by its traffic load. The distribution of exhaust gas, i.e. polluted area and inhabitants affected by the exhaust gas, can be usually modelled and measured following the principle of mathematical morphology. Regarding the exhaust emission as diffusion, the two most fundamental morphologic operations, *dilation* and *erosion* can be applied (see Figure 7), where *A* is the original map feature, *B* is the structuring element, and the blue cell indicates the origin of structuring element and gives the position of the result.



(a) Dilation



(b) Erosion

Figure 7. The fundamental morphologic operations

In our case, *A* represents a road. Here we assume the exhaust spreads evenly and the attenuation is not considered in its effective influence distance. The task is specified as *B*. The influence distance *d* of the exhaust emission depends on the traffic volume and is derived from it. *d* can be then converted into the number of grid cells *b* according to

$$b = integer\ (d/g),$$

where *g* is the size of the grid cell. *B* can now be determined according to *b*. Figure 8 demonstrates the principle of dilation for road emission when *b* is set to 2.

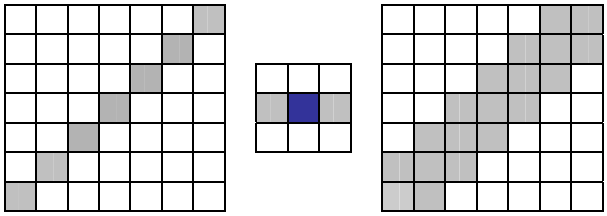


Figure 8. The principle of dilation for road emission

Alternatively we can use the other morphologic operation *erosion*, regarding the road as structuring element, and residential areas as map feature. It yields approximately the same result as with dilation.

3.3.2 Seamless noise indication

Noise is usually measured at discrete points or stations. By spatial interpolation of these points, it is possible to create an isoline map of noise which gives an impression of noise spreading [Stoter 1999]. However, the actual noise level at each particular point relies on a great number of influence factors [Fitzke 1996, 1997].

The *dilation* operation for exhaust emission is not applicable to road noise either for the following reasons: Firstly, the propagation rules of sound wave, especially the acoustic reflection and acoustic attenuation, must be considered. For instance the high buildings can partly reflect the sound wave and more or less shelter from its propagation. Secondly, the *dilation* operation for exhaust emission works well in two-dimensions, but noise is a variable of three spatial dimensions and time. Thirdly, there is a non-linear relationship between the physically measurable noise level and the psychologically perceived level of annoyance, which is partly the reason of country-specific legislation and models of noise pollution.

The environmental noise is originated from roads, railways, aircrafts and industries (e.g. power plant, building site). They are generally distinguished in point, line and area sound sources. If the dimensions of a noise source are small compared with the distance to the receiver, it is called a point source. The sound energy spreads out spherically, so that the sound pressure level is the same for all points at the same distance from the source, and decreases by 6 dB when the distance is doubled.

A line sound source is typically narrow in one direction and long in the other compared to the distance to the receiver. It can be a single source or composed of many point sources operating simultaneously, such as a stream of vehicles on a busy road. The sound level spreads out cylindrically, so the sound pressure level is the same at all points at the same distance from the line, and decreases by 3 dB when the distance is doubled.

The traffic noise is normally modelled as line sound source. When the width of a road is large enough in comparison with the distance from the receiver point to the road, it is regarded as an area source. For area sources, there is theoretically no decrease in intensity with increasing distances.

The important factors for noise propagation include the type of sound source, distance from source, atmospheric absorption, wind strength, temperature and temperature gradient, obstacles such as barriers and buildings, ground absorption, reflections, humidity and precipitation. The contribution of a single sound to the noise can be expressed as follows:

$$L_p = L_w - \sum_{i=0}^{7} \Delta L_i$$

where

L_p Noise level at receiver point in dB
L_w Sound power of the source in dB;
ΔL_1 Correction for attenuation with distance;
ΔL_2 Correction for atmospheric sound absorption;
ΔL_3 Attenuating effect of the walls of buildings;
ΔL_4 Attenuating effect of the outdoor buildings;
ΔL_5 Attenuating effect of the vegetation;
ΔL_6 Correction for barrier attenuation;
ΔL_7 Correction for ground effect (topographical);

The physically measured values at selected discrete points can be used as reference to further adjust the calculated L_p.

The sound power of traffic noise depends on the standard traffic intensity (standard hourly traffic rate in vehicles/h), car and truck portion, permissible maximum speed for car and truck, kind of the road surface, gradient of the road segment, light signal systems. This can be computed by the following formula:

$$L_w = 10 \lg MSV + \sum_{i}^{5} C_i$$

where MSV is the standard traffic intensity and C_1 - C_5 are the corrections for car and truck (heavy traffic) portion, the standard speed, kind of the road surface, gradient of the road section, and the influence of crossings.

Attenuation of the sound spreading with a distance from r_1 to r_2 can be calculated according to following formula:

$$\Delta L = 20 \lg \frac{r_2}{r_1}$$

The absorption of sound by the atmosphere can be calculated according to ANSI 126, ISO3891 or ISO 9613. ISO 9613 Part 1 describes the calculation method for absorption of sound by the atmosphere. For pure tones the standard specifies the attenuation coefficient as a function of frequency, temperature, humidity and pressure.

When a sound source meets obstacles such as buildings, vegetation or ground, it can be reflected, refracted, transmitted and / or diffracted. The results will be the screening and absorption effect which can also attenuate the sound pressure. To calculate this attenuation, the figures and altitudes of buildings or barriers, the distribution and altitudes of the vegetation and the longitudinal inclination of the ground must be considered.

According to the independency principle, the contribution of all sound sources at receiver point can be computed by summing up the separately calculated influences of individual sound sources at the receiver point:

$$L_p = 10 \lg(\sum_i (10^{L_{pi}/10})$$

We use a conic model shown in Figure 9 to compute the physical influences of the sound field on an arbitrary receiver point. By systematically scanning the three dimensional environment around a given receiver point, one or many conic regions can be detected, with each being delimited by a sound source and obstacles between the sound source and the receiver point. The noise propagation from each individual sound source will be then computed and aggregated.

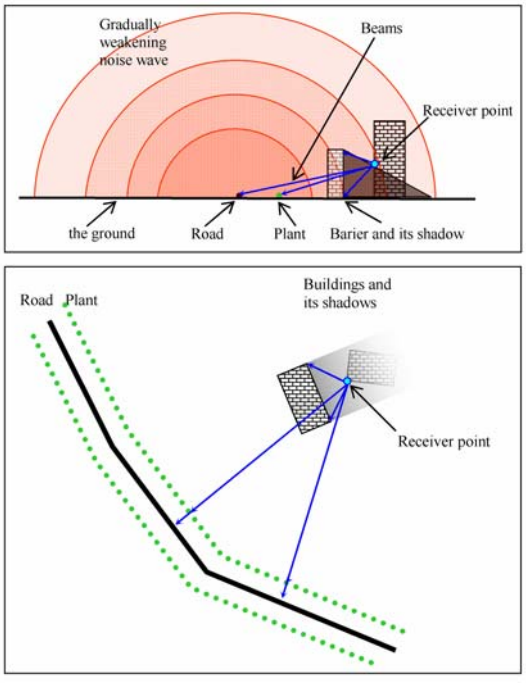


Figure 9. Conic model of sound field

To obtain a seamless noise indication for a given region we use the grid method to systematically divide the region into regular grids and calculate the influence of sound sources on each grid. Based on grid values, a fine-grained noise surface model can be determined. The grid dataset can then be integrated with the Basis-DLM using the method explained in 3.2.

3.4 Integration of the population data

3.4.1 Demographic data

Since the demographic data contains the space-reference, they can be easily bundled with the basic spatial data by the postal data.

However, the integration is based on the post address, e.g. street name and house number. Because a house number has only a pair of coordinates (represents a point), and is often corresponding to a building in which many inhabitants live normally. Therefore, the inhabitants, who correspond to the house number, should be assigned evenly into the area covered by the building.

3.4.2 Integration of commuter data

Commuter data is a special dataset. The number of employees, the number of employees liable to social insurance, and the number of employed commuters liable to social insurance as well as their spatial distributions are known. The overall number of commuters and their distribution over different communes can be estimated by extrapolation. However, a distribution of commuter at address level is more interesting for knowledge discovery. Under the assumption that each house number corresponds to a building or an area and there is a constant commuter increment on each address, we can assign the commuter increment to each inhabitable address as follows:

$$N_{ci} = \frac{N_i}{N_s} \cdot N_c$$

where N_s = the total number of inhabitants
N_i = the number of the inhabitants in a house
N_c = the total number of commuters

4. OUTLOOK

Integrating various datasets from different domains with a fine-grained digital landscape model is a time-consuming but necessary preparation for the subsequent knowledge discovery and data mining. The introduced concepts have been partly implemented with the available test datasets. At the same time of refinement and implementation of integration algorithms, we intend to assure the quality of integration results through comparison with discrete field measurements.

References from Journals:

Sheikholeslami, G., Chatterjee, S. Zhang, A. 2000. WaveCluster: a wavelet-based clustering approach for spatial data in very large databases. *The VLDB Journal 8 (3-4)*, pp. 289-304.

Goodchild, M.F., Siu-Ngan Lam N. 1980. Areal Interpolation: a variant of the traditional spatial problem. *Geo-processing 1(3)*:297-312.

References from Books:

Shekhar, S., 2003. *Spatial Databases: A Tour*. Pearson Eduction, Inc., New Jersey, pp. 321-332.

Yang, Hongguang, 1992. *Zur Integration von Vektor- und Rasterdaten in Geo-Informationssystemen*. Bayerischen Akademie der Wissenschaften, Munich.

References from Other Literature:

Fitzke, J. 1996. GIS-gestützte Berechnung von Schallimmissionen. www.uni-klu.ac.at/groups/geo/gismosim/paper/fitzke/fitzke.htm

Fitzke, J. 1997. Entwicklung eines GIS-Prototyps zur Quantifizierung von Straßenlärmbelastung auf der Grundlage von Schallimmissions- und Einwohnerstrukturdaten. *Salzburger Geographische Materialien*, Band 26, 59-66.

Han, J., Koperski, K. and Stefanovic, N. 1997. GeoMiner: A System Prototype for Spatial Data Mining. In: *Proc. ACM-SIGMOD Int'l Conf. on Management of Data*, Tucson, Arizona, pp. 553-556.

Stoter, J. 1999: Noise Prediction Models and Geographic Information Systems, a sound combination. SIRC 99 – *The 11th Annual Colloquium of the Spatial Information Research Centre*, University of Otago, Dunedin, New Zealand

Wang W., Yang J., and Muntz R., 1997. STING: a statistical information grid approach to spatial data mining. In: *Proc. 23rd Int'l Conf. Very Large Databases*, Morgan Kaufmann, pp. 186-195.

Zhang, M., Shi, W. & Meng, L., 2005: A generic matching algorithm for line networks of different resolutions. In: *Proc. Workshop of ICA Commission on Generalization and Multiple Representation*. Computering Faculty of A Coruña University - Campus de Elviña.