EuroCarto Special Issue

Exploration and Refinement of Regression Tree Models with Interactive Maps and Spatial Data Transformations

Gennady Andrienko*, Natalia Andrienko*, Alexander Ryumkin**, Valery Ryumkin**, Gennady Kravchenko**, Evegeny Tyabaev**, Dmitry Khloptsov**, Svetlana Trofimova**

* Fraunhofer Institute IAIS, Sankt Augustin, Germany, and
City University London, UK
** Tomsk State University, Tomsk, Russia

Abstract. The problem we address is prediction of expected values of some attribute of spatial objects based on values of other attributes, including the geographic positions. A common approach to obtaining such predictions is regression modelling. It is highly desirable that predictive models are not only accurate but also understandable to the users, which gives preference to simpler models. We propose a set of visualization techniques and interactive operations that supports exploration, evaluation, refinement, and simplification of regression tree models. In particular, the analyst can investigate how the model components and their properties are related to the spatial distribution of the objects, and can make the model better account for the spatial aspect of the data by generating new space-based attributes and supplying them to the model building tool.

Keywords. Predictive modelling, geovisualisation, analytical cartography, visual analytics

1. Introduction and related work

A very common task in many domains is prediction of expected values of some attributes based on values of other attributes. Regression models, which try to capture the interdependencies between attributes reflected in available data, are widely used for obtaining such predictions. Besides the prediction accuracy, a highly desirable feature of a model is its under-

standability to the user, which, in turn, requires the model to be sufficiently simple.

The task of attribute prediction may apply, in particular, to attributes of spatial objects, i.e., objects located in geographic space. For spatial objects, not only thematic attributes may be relevant to the prediction task but also the geographic positions. For geographically referenced multi-attribute data, geographically weighted regression modelling has been proposed (Brunsdon et al. 1996). This approach is primarily suitable for phenomena that are spatially continuous (i.e., values exist everywhere in space) and, moreover, spatially smooth (i.e., values in neighbouring locations do not differ much). It is less suitable for discrete spatial objects and for spatially abrupt phenomena. In the latter cases, it may be more appropriate to use the generic regression, which may account for attributes representing the object positions and properties of their spatial distribution. Appropriate attributes may not be originally available in data but may need to be created by the analyst who builds a model.

The work of analysts on building regression models needs to be supported by tools enabling exploration of model quality and other properties as well as model refinement and simplification. Supporting model building, exploration, and refinement by interactive visual interfaces is an actively researched topic in visual analytics. In many works, interactive visualisations are designed to help users to explore, understand, and evaluate a previously built model but not to build or modify the model. Thus, Demšar et al. (2008) employ coordinated linked views and clustering for exploration of a geographically weighted regression model of a spatio-temporal phenomenon. Matković et al. (2010) support users in exploring multiple runs of a simulation model. Visualisation and interaction can reduce the overall number of simulation runs by allowing the user to focus on interesting cases. Maciejewski et al. (2011) suggest an interactive visual interface allowing the user to explore the results of a pandemic simulation model and investigate the impact of various possible decision measures on the course of the pandemic.

Among the works where interactive visual techniques support the process of model building, several papers focus on classification models. Garg et al. (2008) suggest a framework where a classifier is built by means of machine learning methods on the basis of positive and negative examples (patterns) provided by the user through an interactive visual interface; the user finds the patterns using visualisations. Migut and Worring (2010) visualise a classification model, particularly, the decision boundaries between classes. Interactive techniques allow the user to update the model for achieving desired performance. Guo et al. (2009) suggest techniques that help an analyst to discover single and multiple linear trends in multivariate data. Andrienko & Andrienko (2013) propose a framework that supports predictive

modelling of time series of event counts for a discrete tessellation of a territory. There are specific visual analytics methods supporting exploration of relationships between pairs of variables and enabling incremental construction of particular kinds of predictive models, such as linear regression (Zhao et al. 2014) and regression trees (Muehlbacher & Piringer 2013).

In the current paper, we propose a set of visualisations and interactive operations intended to support building, exploration, and refinement of regression models for prediction of thematic attributes of spatial objects. Interactive maps are used for exploring whether and how the prediction accuracy, model components, and model properties are related to the geographic space. The maps are combined with other types of display providing additional perspectives onto model properties and outcomes. Interactive operations allow generation of new attributes based on the spatial distribution of the objects, to be used for progressive model refinement. The overall set of techniques support an iterative process of model building, in which an initial model obtained from an automatic model generation tool is gradually improved due to better accounting for the spatial aspect of the data.

The proposed set of techniques is implemented as a general framework utilizing a bi-directional link between a general purpose open source data mining software Weka (Witten et al. 2011), which includes, in particular, a set of regression modelling algorithms, and a space-time visual analytics research prototype V-Analytics (Andrienko & Andrienko 2006, Andrienko et al. 2013). We demonstrate the use of the techniques on a real world data set of real estate properties. The purpose of modelling is prediction of the relative prices (per square metre) of apartments registered at a real estate agency in Tomsk (Russia) based on historical data.

2. Domain problem settings and example dataset

Real estate business is an extremely important part of the world economy, providing an essential tax share to budgets of local communities (Musgrave & Musgrave 2009, Weingast 2009). A variety of real estate services are developing, including pre-sell object valuation (Adair et al. 1996). Real estate information is used for many purposes, including regional planning at large and estimation of different kinds of payments and compensations. While in many countries planning practices are well established (Davies et al, 1989), specific non-standard procedures can be observed in countries with transition markets (Ryumkin 2006, Qian et al. 1996).

An important aspect of real estate information is price assessment based on object attributes such as location, size, condition etc. A large number of relevant publications is available (see, for example, Kauko & d'Amato 2008). The known approaches are based on mathematical modelling, sta-

tistics (Hui et al. 2010), neural networks, spatial analysis (Anselin 1998), just to name a few. Still, professional assessors are often quite critical to the results of these approaches. While it is not the goal of our work to outperform any of the existing methods of real estate price assessment, we show, by example of regression modelling, how interactive visualisations and data transformations can be used for improving model accuracy and at the same time for model simplification, which improves understandability.

Our example dataset describes about 1,100 objects valuated by a real estate agency in Tomsk (Russia) over a 30 months period from January 2012 till July 2014. Each object is described by the following attributes:

- date of evaluation
- location (longitude and latitude)
- size (in sq.m.)
- price (in RUB)
- material of walls (0=panel building, 1=brick)
- condition (0=rough finish, 1=bad, 2=average, 3=good)

Based on initial visual exploration of the data, some records have been removed because of missing values or obvious errors, e.g., unrealistic values or combinations of values, such as extremely expensive very small apartments. For the remaining 1,070 objects, additional attributes have been derived:

- relative price, RUB per sq.m.
- distance to the city centre (km)

Attribute "condition" has been transformed to 4 binary attributes: Condition_no, Condition_1, Condition_2, and Condition_3.

For building predictive models capturing the dependency of the relative price on the other attributes, we use classification algorithms available in the data mining library Weka (Witten et al. 2011). One of the commonly used approaches to dealing with multi-attribute dependencies is linear regression (Seal 1967). A linear regression model estimates the value of a dependent variable as a linear combination of values of independent variables.

In our example, an application of the linear regression method with 10-fold cross-validation results in the following model:

$$predictedPriceM^2 = 2828915.5193 - 9255.4838*Long - 35205.0744*Lat - 943.2268*Dist - 54.4402*Size - 1057.0997*Condition\_no - 1333.9755*Condition\_1 + 651.4758*Condition\_2 + 2071.8255*Condition\_3 + 2296.9453*Material$$

This model appears quite reasonable as it indicates that the price is influenced positively by the quality (condition) and the material and negatively by the distance from the city centre and the object size (the price per area unit of larger objects is smaller). The regression equation also indicates a geographic trend of price decrease towards the east and north. It should be noted that the independent attributes have different value ranges; therefore, direct comparison of regression coefficients is not meaningful.

While the model corresponds to what could be expected, its accuracy is very low. The correlation between the recorded and predicted prices is rather weak (coefficient=0.305). A more sophisticated modelling method is regression tree (Wang & Witten 1997). In our example, it achieves a slightly better prediction quality with the correlation coefficient of 0.588. However, the model is much more difficult to interpret, as it consists of 21 linear models built for subsets of the data. The division into subsets is done automatically by the algorithm and represented in the form of splitting tree, a fragment of which is shown in Figure 1. Each non-terminal node contains a condition on the values of some independent variable and each terminal node corresponds to one partial linear regression model.

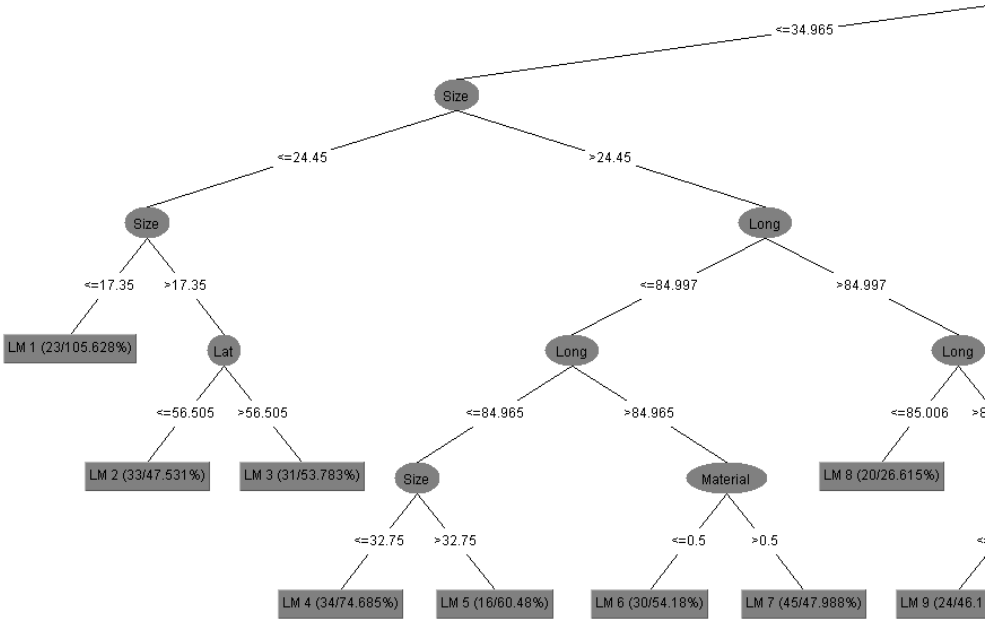

Figure 1. A fragment of a splitting tree showing how a data set is divided into subsets described by different linear regression models.

Since the model accuracy is too low while the complexity is too high, it is appropriate to try to refine and at the same time simplify the model. We shall do this using a procedure of progressive model refinement, which follows the ideas of progressive clustering (Rinzivillo et al. 2008). Before

demonstrating the procedure, we shall introduce the visualizations supporting exploration of model quality and other properties.

3. Visualisation support to model exploration

Exploration of a regression tree model includes the following tasks:

1. Analyse model quality in terms of prediction errors.
2. Analyse model structure in terms of object set partitioning.
3. Analyse the quality of the model components (sub-models) in terms of prediction errors.
4. Analyse the sub-models in terms of the impacts of different attributes on the predicted value.

Task 1. The model quality is assessed based on the statistical and spatial distributions of model errors, i.e., the deviations of the predicted attribute values from the actual ones. From the model building algorithm, the predicted values are obtained. Subtracting the actual values from then gives the absolute errors, from which, in turn, relative errors can be computed as ratios or percentages of the absolute errors to the actual values. The statistical distribution of the errors can be investigated using statistical graphics, such as frequency histograms and box-and-whiskers plot (Fig. 2).

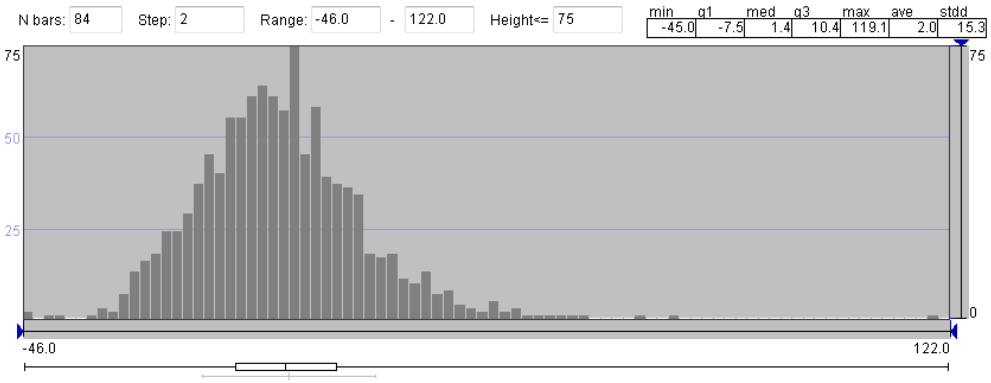

Figure 2. The statistical distribution of relative prediction errors is represented by a frequency histogram, a box-and-whiskers plot below it, and a set of numeric statistical measures in the upper right corner.

The spatial distribution is studied using cartographic visualisations. A possible visualization is demonstrated in Fig.3 (left). The relative errors are represented by proportionally sized circle symbols placed on a map according to the positions of the geographic objects for which the prediction has been done. Two different hues, such as red and blue, are used for representing positive and negative errors, i.e., over-estimation and under-estimation

of the attribute values. Errors can also be represented by colouring or shading; however, in case of point objects, as in our example, representation by symbols or diagrams may be more effective. Interactive filtering allows the analyst to focus on the spatial distribution of high errors, i.e., those that lie beyond a chosen tolerance interval. In our example, the interval is from -20 to 20%.

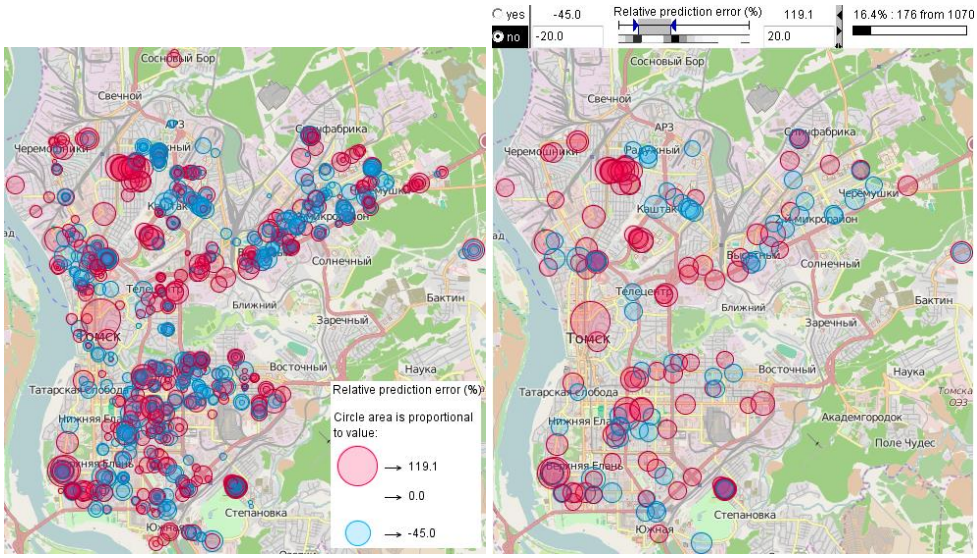

Figure 3. Left: The spatial distribution of the relative prediction errors is shown on a map by circle symbols with sizes proportional to the amount of error and colours showing the sign (positive or negative). Right: As a result of interactive filtering (top), the map shows only the locations where the amount of error is beyond a specified tolerance interval (from -20 to 20% in this example).

Task 2. As explained earlier, a regression tree model consists of several regression models, each covering a subset of objects. The subsets are defined based on values of some attributes chosen by the algorithm. The analyst may be interested to see what subset of objects is covered by each sub-model and whether the division into the subsets has any relation to the spatial distribution of the objects. The modelling algorithm does not directly tell which sub-model corresponds to each object; however, this information can be reconstructed by re-partitioning the object set using the same splitting conditions as in the regression tree. The so obtained sub-model references for the objects are added to the data as values of a new qualitative attribute, which can be visualised on maps using suitable visualisation techniques.

For example, in Fig. 4, left, the sub-model references of individual objects are represented by coloured dots, each unique colour corresponding to one of the sub-models. Using interactive filtering controls, e.g., as shown in the

lower right corner of the map, the analyst can separately view the spatial distribution of each object subset or compare the distributions of two subsets. To see more explicitly the areas in space covered by the object subsets, the analyst may build and visualise the spatial convex hulls of the subsets. An example is shown in Fig. 4, right. The hull contours are semi-transparently filled with the colours assigned to the sub-models. It can be seen that the areas covered by sub-models 2 and 3 almost coincide, and the same is true for sub-models 4 and 5. These two pairs of sub-models are spatially separated, while the area of sub-model 1 extends almost over the whole territory and almost completely covers the areas of the other sub-models.

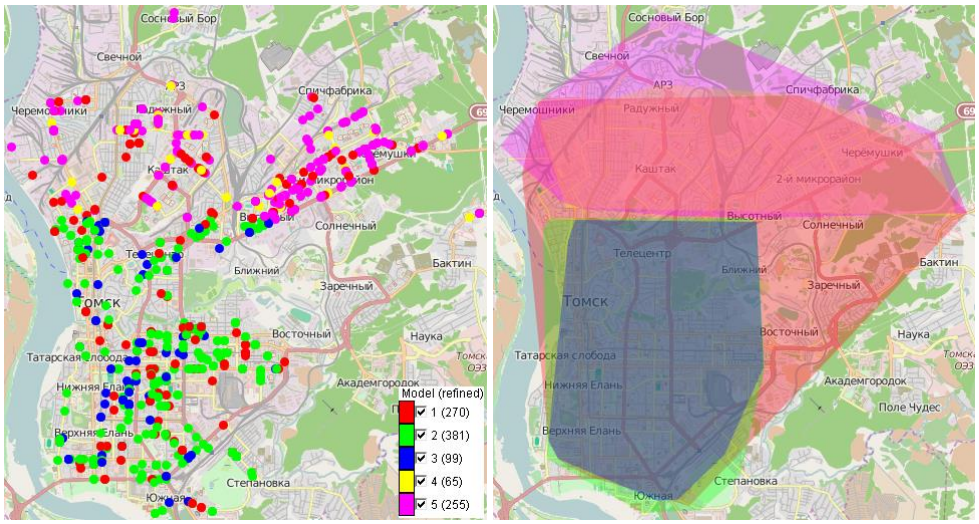

Figure 4. Left: Each object is represented by a dot coloured according to the regression tree sub-model applicable to this object. Right: the areas in space covered by the object subsets are shown in the form of their spatial convex hulls.

Task 3. The interactive legend shown in the lower right corner of the map in Fig. 4, left, and used for selecting object subsets for viewing, affects not only the map it belongs to but also all other displays currently present on the screen. In particular, if there exists another map showing the prediction errors, as in Fig. 3, this map will only show the errors for the currently selected object subset(s), i.e., the subset selection works as a filter. Similarly, the statistical displays will represent the statistical distribution of the values for the currently selected object subset(s). Hence, the analyst can conveniently analyse the spatial and statistical distributions of the prediction errors for each sub-model. Moreover, the object subset filter is automatically combined with all other filters that are currently in operation, such as the attribute-based filter shown on the top right of Fig. 3 (the filter selects the error values beyond the tolerance interval from -20 to 20%). This allows the

analyst to specifically look at the spatial distribution of high prediction errors for each sub-model, as demonstrated in Fig. 5.

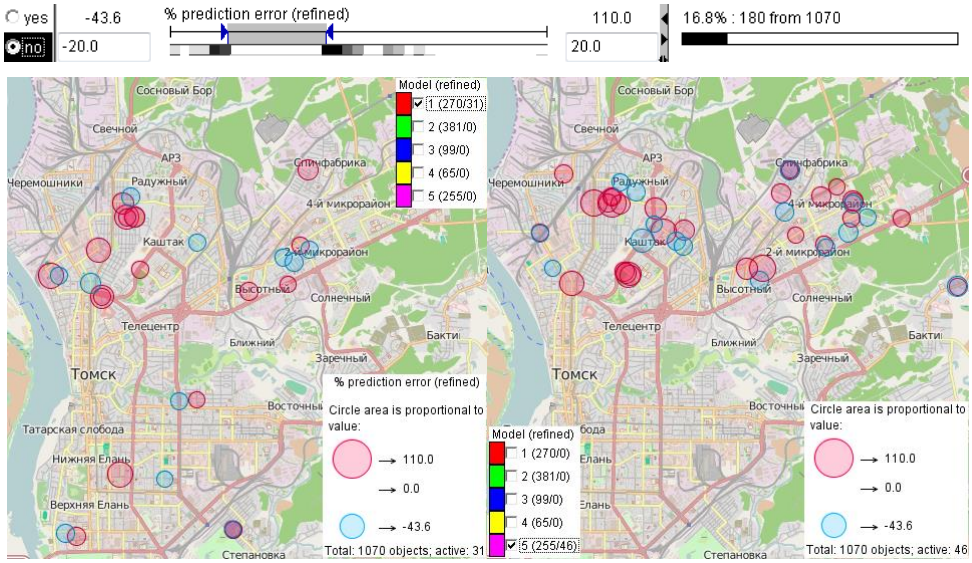

Figure 5. By interactive selection of object subsets dealt with by different sub-models, the analyst can investigate the spatial distribution of the prediction errors separately for each sub-model. The two screenshots refer to sub-models 1 (left) and 5 (right). The interactive legend enabling subset selection is shown in the upper right corner of the left map and in the lower left corner of the right map. The subset selection is combined with the attribute-based filter shown above the maps.

Task 4. The impact of different attributes on the prediction can be judged from the corresponding coefficients in the regression formulas of the sub-models. Since the attributes used for the prediction may not be comparable, as in our example, the coefficients for different attributes are also not directly comparable. However, the coefficients for the same attribute in different sub-models are comparable. For a convenient assessment and comparison of the relative impacts of different attributes, the coefficients need to be brought to a common range, such as from -1 (highest negative impact) to 1 (highest positive impact) or from -100 to +100%. This may be done by dividing each coefficient in each sub-model by the maximal modulus (i.e., absolute value) among the coefficients for the same attribute in all sub-models. To obtain percentages, the ratios are multiplied by 100.

The original or transformed coefficients from the different sub-models can be visualised on multi-attribute displays, such as the bar chart shown in Fig. 6, where the rows correspond to the different attributes and each group of bars of the same colour corresponds to one of the sub-models. The bar lengths are proportional to the absolute values of the transformed coeffi-

cients, and the bar orientation (left or right) shows their signs, which, in turn, show whether the impacts are negative or positive.

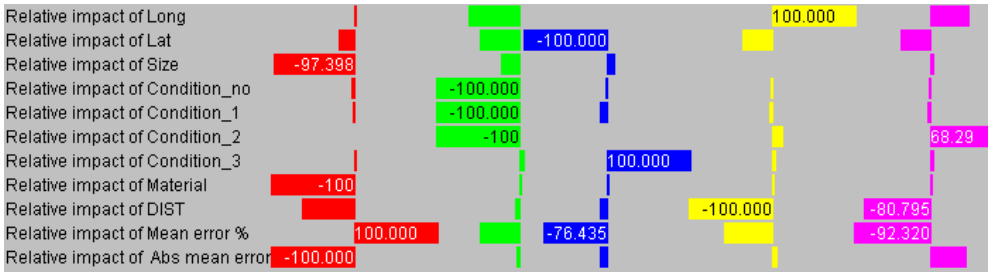

Figure 6. A bar chart display enables comparison of the impacts of different attributes on the prediction in different sub-models.

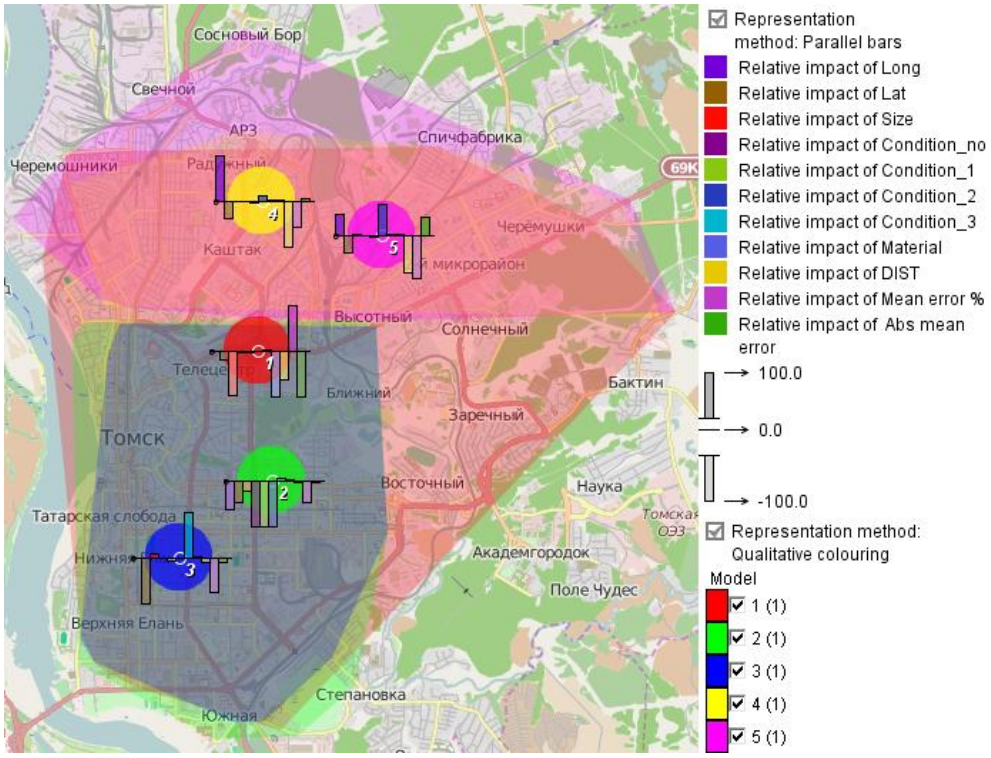

Figure 7. The attribute impacts are visualised on a diagram map. Each diagram corresponds to one sub-model. The bar heights are proportional to the amounts of impact and the bar orientations (up or down) to the signs (positive or negative).

For relating the attribute impact indicators, i.e., the transformed coefficients, to the spatial context, the indicator values can be visualised by diagrams drawn on a map. The problem is that the coverage areas of the sub-models may greatly overlap or even nearly coincide. When a diagram is put at the centre of each area, the diagrams corresponding to different sub-

models may overlap, which complicates reading. Therefore, it is reasonable to select a representative point on a map for each sub-model so that the points are sufficiently distant from each other, and use these points as the positions for the diagrams. The points can be interactively specified by the analyst, e.g., by mouse-pointing on the map. Thus, the analyst may select representative points within spatial concentrations of objects belonging to the different subsets.

In Fig. 7, the relative impacts of the different attributes in the sub-models are visualised using bar diagrams, which are placed on a map at the positions of selected representative points for the sub-models. In one diagram, each bar represents the impact of one attribute. The bar lengths are proportional to the amounts of impact and the orientations (up or down) show the impact signs, i.e., positive or negative. To facilitate the association of the diagrams to the sub-models, which are consistently represented by distinct colours in various displays, the diagrams are drawn on top of circles painted in the colours assigned to the models.

An alternative visualisation of the same information is "small multiples" (Tufte 2001, 1983), i.e., a display with multiple small maps, each representing the relative impact indicators for one attribute in the different models. An example is demonstrated in Fig. 8. In each small map, the relative impact values for one attribute are represented by standalone bars with the heights proportional to the amounts of impact. The impact signs (positive or negative) are represented by bar orientations (up or down) and, additionally, by bar colours (orange or cyan).

The diagram map in Fig. 7 allows comparison of the overall profiles of the impacts of the different attributes among the sub-models. Thus, if the sub-model coverage areas differ, the analyst may look whether sub-models covering close or overlapping areas have similar attribute impact profiles. In our example, this is the case for the pair of sub-models 4 and 5, the overlapping coverage areas of which are located on the north, and not the case for the sub-model pair 2 and 3, the coverage areas of which are almost identical but the profiles are quite different.

The small multiple map in Fig. 8 allows the analyst to consider each attribute separately and compare the sub-models with regard to the impacts of each attribute. In particular, the analyst can judge for each attribute whether the differences between the impact values are related to the spatial positions of the objects. Thus, in our example, the attribute "Condition_2" has positive impacts on the north of the territory and negative or no impact on the south. It can also be seen that model 1 (red), which covers the whole territory, strikingly differs, in terms of the attribute impacts, from the other models, which cover the northern or southern parts of the territory.

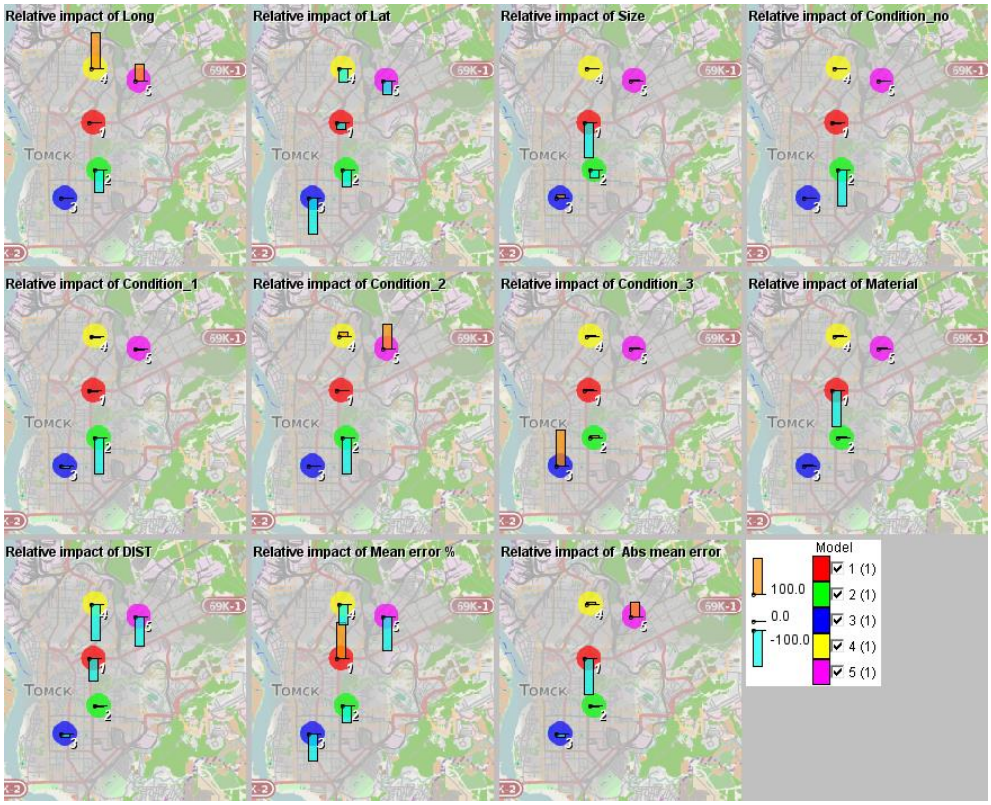

Figure 8. In a "small multiple" map display, each small map shows the relative impacts of one attribute on the prediction in different sub-models. The amount of impact is represented by bar heights and the sign of impact (positive or negative) by bar orientation (up or down) and colour (orange or cyan).

In Figs. 4 to 8, we have presented visualisations that can be used for the exploration of model components (sub-models). However, all these methods can be effective only in the case of a relatively small number of components. Even irrespectively of the visualisation techniques, it is very tedious and time consuming to explore a large number of components, for example, the 21 sub-models of the tree a fragment of which is shown in Fig. 1. Such an activity may be even meaningless because a complex model with a large number of components is not necessarily more accurate as a simpler model. Therefore, it is reasonable to try to simplify and, if possible, simultaneously refine the model before starting the exploration of its components. Thus, the illustrations in Figs. 4 to 8 have been produced using an already simplified model. In the next section, we propose a set of visually supported interactive operations that support model simplification and refinement.

4. Interactive operations for model simplification and refinement

Since there is no way to intervene in the work of the model building algorithm, the only way to change its output is by modifying the input. There are, obviously, two basic ways to modify the input: change the set of objects and change the set of attributes. In the context of our work, we are specifically interested in changes that make the algorithm better account for the spatial aspect of the data, i.e., for the object distribution in space. The existing classification algorithms can account for the spatial aspect only if it is represented by some attributes along with the other (thematic) attributes to be used for the prediction. In particular, the geographic coordinates (longitudes and latitudes) of the spatial objects can be supplied to a model builder as attributes, and they will be treated as usual numeric attributes. So we did in our example, where the coordinates are represented by attributes "Long" (longitude) and "Lat" (latitude). The coordinate-based numeric attributes may participate in the object set partitioning and/or in the regression formulas of the sub-models. However, this basic way of accounting for the spatial aspect may be insufficient for obtaining a good (i.e., sufficiently accurate while simple) prediction model. In this section, we shall consider possible additional ways for involving the spatial aspect in the modelling.

4.1. Modification of the object set

After obtaining an initial model, it is reasonable to look whether and how the spatial aspect (so far only in the form of object coordinates) is represented in this model. However, the initial model may be too complex for exploration, as in our example, where the model has 21 sub-models. To simplify a too complex initial model and thereby make it better suitable for interactive exploration, we propose to temporarily remove "prediction quality outliers" from the object set, i.e., the objects for which the prediction errors of the initial model are very high.

1.1.1. Removing outliers

The outliers can be identified based on the statistical distribution of the model errors, for example, by applying the definition of an outlier adopted in statistics: an outlier is a value lying beyond the interval [median – 1.5*IQR, median + 1.5*IQR], where IQR is the inter-quartile range, i.e., the difference between the third and the first quartiles. The reduced object set is then sent to the model building tool, which can be expected to produce a simpler model.

In our example (see the statistics on the top right of Fig. 2), the median of the relative prediction errors is 1.4, the first and third quartiles are -7.5 and

10.4, respectively, the interquartile range is thus 17.9, and the interval to be used for identifying the outliers is [-25.45, 28.25]. We use the attribute-based filter to select only the objects with the relative error values within this interval and remove the outliers. The "cleaned" object set consists of 996 objects out of 1,070, i.e., 74 objects (6.9% of all) are outliers. We send the reduced object set to the model building tool, which produces a model having only 8 sub-models (see Fig. 9, left) and a much better correlation (0.68) between the predicted and original price values than in the original model (0.588). Now we can more conveniently investigate whether the object coordinates have been taken into account in the object set partitioning by the model builder. We see that the longitude has played some role but rather moderate since it was used for making quite small object subsets (with 25, 41, and 44 objects). The part of the tree in which the division by the longitudes is done is marked in Fig. 9, left (it is in the lower right corner of the tree view). On the right of Fig. 9, the spatial distribution of three object subsets corresponding to this part of the tree is shown on a map; the dot colours correspond to the three different sub-models residing in the tree leaves.

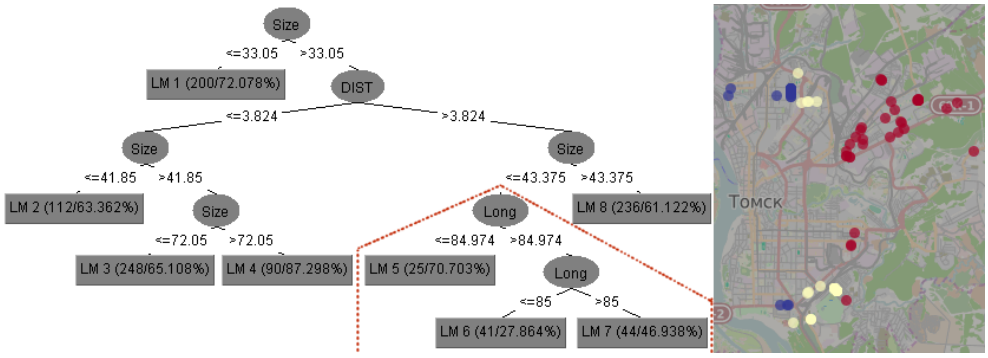

Figure 9. Left: The splitting tree shows the partitioning of the object set in a simplified model resulting from the outlier removal. The red dotted polygonal line marks the part of the tree where the geographic coordinates (longitudes) are used in defining object subsets. Right: The map shows the spatial distribution of the object subsets corresponding to the marked part of the tree. The dot colours represent the three sub-models located in the leaves of the marked part of the tree.

We see that the division according to the longitude is too crude and, generally, does not respect the natural spatial grouping of the objects. However, all but three objects represented by the red dots make a kind of spatial cluster on the northeast. This part of the city is separated from the rest by a railway line. It may be reasonable to try to apply the modelling algorithm separately to this part and to the remaining part of the territory. For this purpose, the object set can be interactively divided into geography-related subsets using the map interface.

As we mentioned, the outlier removal is used as a temporary measure for reducing the complexity of the initial tree and allowing the analyst to see which attributes have what impact on the object set division. After gaining this understanding, the filter used for outlier removal is cancelled, and the analyst further works with the complete object set.

1.1.2. Interactive subdivision of the object set

A possible interactive interface for geography-based object set division is demonstrated in Fig. 10. On the right, a set of interactive controls allows the analyst to create an arbitrary number of classes, give names (labels) to them, and choose class colours. By mouse clicking or mouse dragging on a map, the analyst can select a group of objects and, by pressing the button "Selection >> class", assign them to one of the classes. In this way, various geography-based object classifications can be created. In the example in Fig. 10, we have divided the set of objects into three classes: "NE" located in the north-eastern part of the city, which is separated from the rest by a railway, "S" located on the south and separated from the rest by a small river, and "N + centre" including the remaining objects. The classes are represented on the map by dot colours: red, blue, and yellow, respectively.

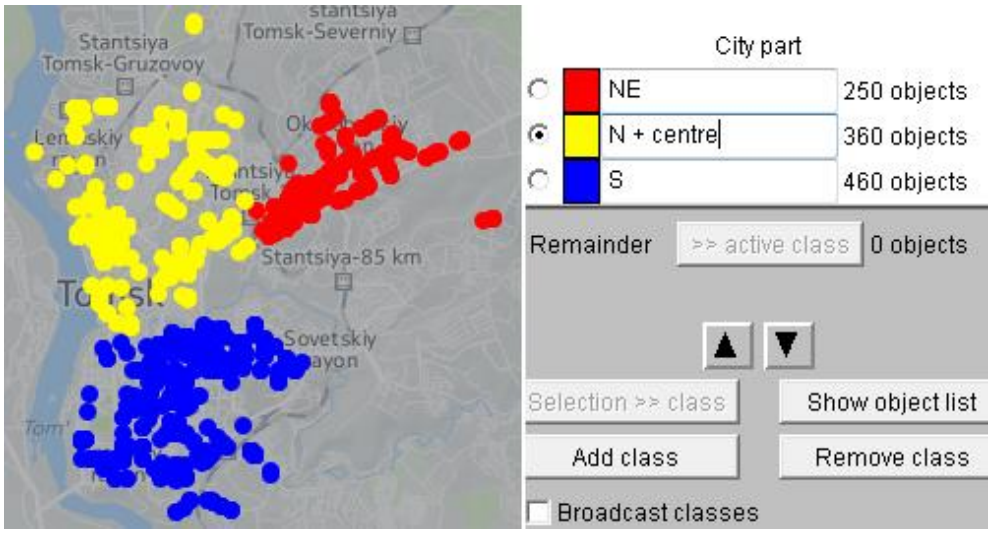

Figure 10. The object set is interactively divided into subsets based on the spatial distribution of the objects and the properties of the underlying territory.

The purpose of the object set partitioning is to try to refine and simplify the regression tree model by applying the modelling algorithm separately to parts of the territory that differ in their properties. Obviously, to achieve simplification, the number of parts to consider should not be large.

In our example, the application of the algorithm separately to the class "NE" results in a regression tree consisting of only three sub-models, the object subset being subdivided based on the property sizes. Along with the simplification, the prediction quality substantially increases: the correlation between the predicted and actual values is 0.645 instead of 0.588 for the initial model. However, the results for the classes "S" and "N + centre" are not as good as for "NE". The trees consist of 5 and 8 sub-models, and the correlation coefficients are 0.456 and 0.569, respectively, i.e., worse than it was initially. Since the results for "NE" are good, it is reasonable to keep this part separate and apply the algorithm to all remaining objects taken together, i.e., to the union of the classes "S" and "N + centre". When we do this, we obtain a regression tree with only 4 sub-models, the object set being also subdivided based on the property sizes, as for the "NE". The correlation coefficient is 0.553, i.e., the prediction quality for the two joint classes is slightly lower than for the initial model. Hence, we have achieved a substantial simplification by replacing the initial 21 sub-models by only 7 (3 for "NE" and 4 for the rest), but we need to work further on improving the prediction quality, especially in the western part of the city.

4.2. Generation of additional space-related attributes

Low prediction quality of a model often means that the target attribute (i.e., the values of which is predicted) is influenced by some factors that are not reflected in the available data. The only way to improve the prediction quality in this case is to obtain additional data that can be attached to the objects as values of new attributes. Additional attributes of objects under study can be derived from other datasets related to the same territory. For example, if we had data concerning the levels of air pollution or noise over the city of Tomsk, we could derive the pollution or noise values at the location of each property and let the modelling algorithm make use of these values. It can be expected that the environment characteristics affect the real estate prices, and accounting for these characteristics in model building can improve the model prediction quality. Unfortunately, we cannot demonstrate this approach to data enrichment since we have no additional data for the territory of Tomsk.

Another (complementary) approach is to take into account the spatial distribution of the model errors. Thus, Fig. 3 suggests that positive and negative prediction errors may tend to be spatially grouped. The idea of the approach is to (1) divide the territory into compartments accounting for the spatial distribution of the objects, (2) compute for each compartment the mean prediction error of the current model, (3) for each object, attach the mean prediction error from the containing compartment as the value of a new attribute, and (4) supply the new attribute values together with the

previously existing ones to the modelling algorithm. New attributes can be generated based on both the absolute and relative prediction errors of the current model. It can be expected that the modelling algorithm will use the new attributes for making corrections in the prediction.

The approach is illustrated in Fig. 11. For partitioning the territory into compartments based on the object distribution, we use the point clustering algorithm proposed by Andrienko (2011). The algorithm organises points into groups fitting in circles with a given maximal radius. The centres of the groups are then taken as generating seeds for the Voronoi tessellation. By varying the radius, larger or smaller spatial compartments can be obtained. On the left of Fig. 11, the compartments have been obtained using the maximal radius of 2000 metres. The red dots are the centres of the point clusters built by the clustering algorithm and simultaneously the generating seeds for the Voronoi polygons. The polygons are shaded according to the computed mean relative errors, with shades of brown representing positive errors and shades of blue negative errors. The errors have been computed based on the predictions obtained from the combination of two regression trees (for the northeast and for the rest of the territory) described in the previous section.

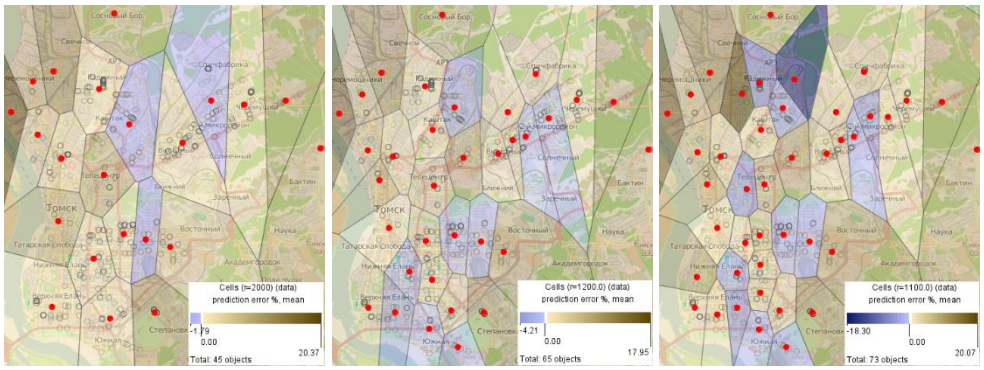

Figure 11. The territory is divided into compartments based on the spatial distribution of the point objects. Summary statistics of the prediction errors are computed for the compartments and used in modelling, along with the original attributes, for correcting the prediction error. This approach can be tried with different compartment sizes.

After building the polygons and computing the mean absolute and relative errors within each polygon, we transfer these computed values to the objects contained in the polygons. Then we apply the modelling tool to the original attributes of the objects plus the two new attributes, i.e., the absolute and relative errors within the containing polygons. The result of the application is encouraging: the model includes only 8 sub-models, and the correlation is slightly higher: 0.603 against 0.588. To further improve the results, we try finer territory divisions by decreasing the maximal allowed

radius of a point cluster for the clustering algorithm. In the centre of Fig. 11, the division has been obtained with the maximal radius of 1200 m. The results for this division do not differ much from the previous one: 8 sub-models with correlation coefficient of 0.614. However, decreasing the maximal radius to 1100 m leads to a regression tree with only 5 sub-models and correlation coefficient of 0.629. The illustrations in Figs. 4 to 8 are based on this result. Interestingly, an attempt to build separate models for the north-eastern parts and for the rest does not really improve the model. The total number of sub-models increases to 10 (3 for the northeast and 7 for the rest). The correlation coefficients are 0.641 for the northeast and 0.607 for the rest. While the latter is better than it was without accounting for the model errors, the former is slightly worse. There is no clear quality gain compared to the variant with 5 sub-models but the complexity is twice as much. Hence, it makes sense to use the simpler model built for the whole object set.

The territory division can be further refined by decreasing the maximal cluster radius for the clustering algorithm. With the maximal radius of 900 m, we obtain a yet simpler tree with only 4 sub-models and correlation coefficient 0.631. It should be borne in mind, however, that too fine division of the territory may lead to model over-fitting, i.e., reproducing occasional fluctuations rather than essential spatial patterns. The analyst should decide what level of division is appropriate based on the knowledge of the territory and the data.

5. Conclusion

As was said at the beginning, it was not the goal of this work to build a perfect model for real estate price prediction. Our goal was to support model understanding, exploration, simplification, and refinement, specifically, taking into account the spatial aspect of data used as input for model building. For the real estate data, which were used as a test case and as an example in this paper, we could achieve great simplification (from 21 sub-models to 5 or even 4) and moderate improvement of the prediction quality. It is clear that the available data do not reflect all factors affecting the prices, and additional data would be necessary for further quality improvement. However, the available data were sufficient for our purposes: we have found and demonstrated the possible ways of making simpler and better models using interactive map-based operations and spatial computations.

The techniques proposed in this paper are suited to a particular kind of predictive models, namely, regression tree classification models. They are applicable to spatially referenced objects, not necessarily points, character-

ised by multiple thematic attributes. The techniques and procedures can thus be used in a variety of application domains. We would like to stress that all proposed techniques involve interactive maps and map-based interfaces for spatial clustering and spatial computations. Effective use of these techniques for modelling is underpinned by a bi-directional link between the visualisation system and the model building tool.

Acknowledgement

This work was partially funded by European Commission within projects SoBigData ("Social Big Data Research Infrastructure", grant agreement 654024) and VaVeL ("Variety, Veracity, VaLue: Handling the Multiplicity of Urban Sensors", grant agreement 688380).

References

Adair A, Downie M, McGreal S (1996) European valuation practice. E&FN Spon, London

Andrienko N, Andrienko G (2006) Exploratory analysis of spatial and temporal data: a systematic approach. Springer, Berlin

Andrienko N., Andrienko G. (2011) Spatial generalization and aggregation of massive movement data, IEEE Trans. Visualization and Computer Graphics, 17(2): 205-219

Andrienko G, Andrienko N, Bak P, Keim D, Wrobel S (2013) Visual analytics of movement. Springer, Berlin

Andrienko N, Andrienko G (2013) A visual analytics framework for spatio-temporal analysis and modelling. Data Mining and Knowledge Discovery 27(1):55–83

Anselin L (1998) GIS Research Infrastructure for Spatial Analysis of Real Estate Markets. Journal of Housing Research 9(1):113–133

Brunsdon C, Fotheringham AS, Charlton ME (1996) Geographically weighted regression: a method for exploring spatial nonstationarity. Geographical analysis 28(4), 281–298

Davies H, Edwards D, Punter J, Hooper A (1989) Planning control in Western Europe. HMSO, London

Demšar, U., Fotheringham A.S., Charlton M. (2008) Exploring the spatio-temporal dynamics of geographical processes with Geographically Weighted Regression and Geovisual Analytics. Information Visualization, 7, pp.181-197

Garg S, Ramakrishnan IV, Mueller KA (2010) Visual Analytics Approach to Model Learning. In Proc. IEEE Symposium on Visual Analytics Science and Technology VAST'10, pp.67-74

Guo,Z., Ward, M.O., Rundensteiner, E.A. (2009) Model Space Visualization for Multivariate Linear Trend Discovery. In Proc. IEEE Symposium on Visual Analytics Science and Technology VAST'09, pp.75-82

Hui SK, Cheung A, Pang JA (2010) Hierarchical Bayesian approach for residential property valuation. International Real Estate Review 13(1):1–29

Kauko T, d'Amato M (2008) Mass appraisal methods: an international perspective for property valuers. Wiley-Blackwell

Maciejewski, R., Livengood, P., Rudolph, S., Collins, T.F., Ebert, D.S., Brigantic, R.T., Corley, C.D., Muller, G.A., Sanders, S.W. (2011) A Pandemic Influenza Modeling and Visualization Tool. Journal of Visual Languages and Computing, 22, pp.268-278

Matković, K., Gračanin, D., Jelović, M., Ammer, A., Lež, A., Hauser, H. (2010) Interactive Visual Analysis of Multiple Simulation Runs Using the Simulation Model View: Understanding and Tuning of an Electronic Unit Injector, IEEE Transactions on Visualization and Computer Graphics, 16(6), pp.1449-1457

Migut, M., Worring, M. (2010) Visual Exploration of Classification Models for Risk Assessment. In Proc. IEEE Symposium on Visual Analytics Science and Technology VAST'10, pp. 11-18

Muhlbacher T, Piringer H (2013) A partition-based framework for building and validating regression models. IEEE Transactions on Visualization and Computer Graphics 19(12):1962–1971

Musgrave R, Musgrave P (2009) Public finance: theory and practice. Business Atlas Publ.

Qian Y, Weingast B (1996) China's transition to markets: market-preserving federalism, Chinese style. Journal of Policy Reform 1:149–185

Rinzivillo S, Pedreschi D, Nanni M, Giannotti F, Andrienko N, Andrienko G (2008) Visually driven analysis of movement data by progressive clustering. Information Visualization 7(3/4):225–239

Ryumkin A (2006) Reforming the management framework of urban land use. Studies on Russian Economic Development 17(2):83-90

Seal HL (1967) The historical development of the Gauss linear model. Biometrika 54(1/2):1–24

Tufte ER (2001) [1983] The Visual Display of Quantitative Information (2nd ed.), Cheshire, CT: Graphics Press

Weingast B (2009) Second generation fiscal federalism: The implications of fiscal incentives. Journal of Urban Economics 65:279–293

Witten IH, Frank E, Hall MA (2011) Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, New York

Wang Y, Witten IH (1997) Induction of model trees for predicting continuous classes. Poster papers of the 9th European Conference on Machine Learning

Zhao K, Ward MO, Rundensteiner EA, Higgings HN (2014) LoVis: local pattern visualization for model refinement. Computer Graphics Forum 33(3):331–340