DE GRUYTER
OPEN

MAPPING DATA - QUALITY, QUANTITY OR BOTH?

N.Sz. Suba [a, *], Șt. Suba [a]

[a] University of Oradea, Faculty of Architecture and Constructions, Department of Cadastre and Architecture, str. B. St. Delavrancea nr. 4, 410058 Oradea, Romania, * e-mail: suba_norbert@yahoo.com

Received: 10.03.2015 / Accepted: 30.03.2015 / Revised: 26.04.2015 / Available online: 31.05.2015

DOI: 10.1515/jaes-2015-0013

KEY WORDS: Mapping Data Quality, Data Sampling, Mapping Data Error Estimation, Quality Standards

ABSTRACT:

In the history of mankind, society was always heavily relying on the knowledge regarding their surroundings. Maps of already known areas were used in different areas of human activity, regardless if they were military applications or for plain orientation purposes. In the modern days of the 21st century, it is almost unimaginable for any corporate or private user to carry out their usual activity without precise knowledge of their (eventually larger than before) surroundings. One thing that changed in the past few years is that although people always used maps, the making of these were in the hand of certain specialists - the land surveyor engineers. Today, to a certain extent, any user can contribute to the expansion of existing maps (or even create new ones), thus leading to a vast dataset included in these maps. While this will theoretically further expand our knowledge regarding our surroundings, a natural question can come in everybody's mind: does the increasing quantity lead to unassailable quality as well? Can we enhance the reliability of existing maps and contribute to nearly error-free future maps?

1. INTRODUCTION

1.1 General Introduction

In the history of mankind, society was always heavily relying on the knowledge regarding their surroundings. Maps of already known areas were used in different areas of human activity, regardless if they were military applications or for plain orientation purposes. In the modern days of the 21st century, it is almost unimaginable for any corporate or private user to carry out their usual activity without precise knowledge of their (eventually larger than before) surroundings.
One thing that changed in the past few years is that although people always used maps, the making of these were in the hand of certain specialists - the land surveyor engineers, although the registering of, for example, the cadastral data was always a public affair (Volkan, 2011). Today, to a certain extent, any user can contribute to the expansion of existing maps (or even create new ones), thus leading to a vast dataset included in these maps. While this will theoretically further expand our knowledge regarding our surroundings, a natural question can come in everybody's mind: does the increasing quantity lead to unassailable quality as well? Or it simply creates a so-called data-silo with further redundant data (Suba, 2014)?

The answer is no, quality will not increase by improving the sheer quantity. Although not all users

Corresponding author.

require pinpoint accuracy in the terms of geodesy (millimetres' precision), maps can contain other data that can still negatively affect any user. The precision of maps is not necessarily expressed in terms of pinpoint positional accuracy of the details contained. Rather than that, it's the details that matter. Outdated or misleading/incomplete information can be as harmful as mispositioned details. Certain users will happily trade the millimetric positional possibilities for an accurate knowledge regarding the metadata in the maps. Certainly, the ultimate goal of mapping would be to provide pinpoint accuracy position of all the details, along with temporal accuracy - but with so many contributing to these databases, will this ever be possible?

1.2 Spatial data quality

The collection of spatial data is influenced by random errors, methodical errors and gross errors.

The quality in mapping applications depends on the quality of stored data. In mapping applications, the data quality has a great influence over the results. The data are used without considering the contained errors, and this can lead to erroneous results, disorienting information and bad decisions that can produce high costs to the user. Any difference between the real world and the dataset is considered an error (Joos, 2006).

There are two key points for enhancing data quality – prevention and correction. Error prevention is considered being highly superior to error detection, because error detection is usually costly and doesn't guarantee a 100% success rate. The best way to prevent errors is to follow the regulations regarding the specifications, structure, dataset, the implementation and evaluation procedures (Suba, 2010a). Implementing existing ISO standards for future expansion of the existing maps will enhance their quality (and thus, their reliability) (van Oosterom, 2006), but there is a need to address the problems of already existing maps as well. Unfortunately, data quality management is rather seen as a cost rather than a way to change things around (Jakobsson, 2009).

The graphical and non-graphical database in case of the digital maps (or derived from them) must satisfy well defined requirements or requirements which appeared from the users' part. Complete verification of the objectives and of the whole dataset is not economically advantageous, neither for the data producer, nor for the user, having in mind the costs of verification, evaluation or validation. A complete verification of the work, having in mind the statute of a verification engineer, can cost as much as twice the cost of obtaining the data in the field.

Carefully choosing the right sampling method can lead to better results, as well as reduced costs, and as we all know, none of these can be overlooked. Moreover, sampling methods are known to perform well in other areas of scientific activities, when analyzing a reduced sample pattern is more efficient than analyzing the whole dataset (Oteros, 2013). The sampling methods presented in this paper will address, at some point, the specific needs of cadastral maps, but the general rules will still apply to any type of analysed spatial dataset. Taking into account the advantages of the integral data acquisition, which comes with less costs and effort (Zahir, 2012), a certain dataset will still need specific check for errors.

The notation „map" is used as an abbreviation in this paper, but it refers both to the maps on paper support at large scales and to the digital map, as well as the (meta)database of the digital maps.

2. MATERIALS AND METHODS

2.1 Data verification principles

As it was mentioned earlier, complete verification of the objectives of the whole dataset is not economically advantageous neither for the data producer, nor for the user, having in mind the costs of verification, evaluation or validation.

In the ISO standards, the work process of taking samples is specified and also sampling schemes of the datasets. In order to reveal the samples for evaluating conformity in accordance to specifications of a product, ISO 2859 and ISO 3951-1 (ISO 3951-1) series can be applied. These standards were initially developed for non-spatial purposes.

Lot and article are important concepts in the specified sampling inspection method of ISO 2859 and ISO 3951-1 series. A lot is the minimal unit for which quality can be evaluated.
An article (object) is the minimal unit that can be inspected, and should be defined by the data producer in conformity with the product specification.

Size of the sample (pattern) - The size of the population, and, consequently, the size of the sample cannot be defined in accordance with the density of the articles (objects). Defining the size of a sample needs explicit indication of the elements.

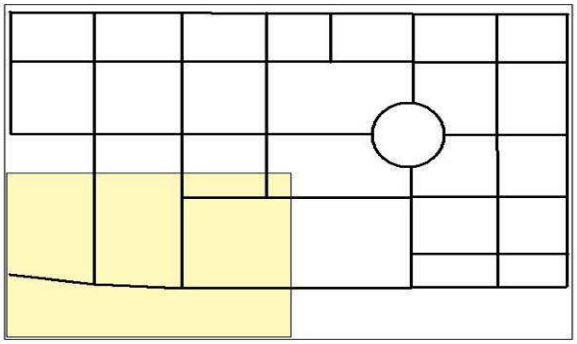

Figure 1. Data volume in a sample

Figure 1 represents data from the application domain of data quality. It describes a possible sample zone, which is 25% from the total area of the application domain, but only 5% of the length of the road network and about 10% of the nodes.

We try to bypass the difficulties of such samplings, by trying to define the size and location of a sample using a combination of various criteria, in order to find the representative samples.

2.2 Sampling strategies

In order to define the samples and the sampling methods, we must take into account the particular aspects of topographic and cadastral data.
There are two aspects which can be taken into account at the sampling strategy:

- articles (elements) which are to be sampled (zone or characteristics);
- the manner in which elements are selected (probability or judgment, decision).

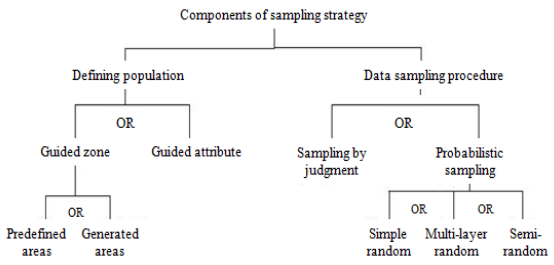

Figure 2. Relations regarding the strategy of sampling

Now to move on to the components of the sampling strategy, we will continue with the sampling methods.

2.2.1 Sampling methods

Guided sampling of properties (non-spatial sampling) - Using this method, samples are selected on non-spatial attributes, and in this way, selection is not dependent on spatial localization. A sample in the application domain of data quality can be randomly selected, assuming homogeneous production characteristics for the entire dataset. In this case, a simple random selection cannot offer the expected result, because there is a chance that these characteristics are not evenly distributed. Surely, a layered sampling or a semi-random sampling would give better results. Semi-random sampling can assure the inspection of various criteria for tightening the sampling, reducing the costs of the associated inspection process.

Guided sampling of the zone (spatial sampling) - The guided zone sampling strategy is based on spatial considerations. Sampling units include geographical domains, social-political domains, and existing administrative domains. Having this in mind, we can also use other partitions from the definition domain for which inspection is required. This type of sampling can be used in a first stage, and it will be followed by an attribute-guided sampling from each zone selected for inspection. The guided sampling can be divided in three big categories:

- domain of equal squares obtained by randomly generating their centre point, with the condition that these squares cannot overlap;
- regular grill of cells, which cover the whole area of the definition domain; in this case, it has to be defined the rate of included or excluded cells, which are not entirely in the interior of the inspected domain. It is used when product specifications require inspection of the whole surface (area). These first two categories are part of the zone generation domain.
- sampling on predefined areas is used when the distribution of characteristics is not homogeneous, spatial partitioning with different sizes and different domains is required for the semi-random sampling.

2.2.2 Sampling Procedures

Sampling by decision (judgment) - For sampling by statistical methods we apply sampling theories, and

this involves random selections of probe elements. The main characteristic of the probabilistic sampling is that each member of the population from where the sample is selected has a known selection. Using probabilistic sampling, we can make statistical presumptions about the sampled population. Sampling models based on decisions presumes the selection of the sampling based on the knowledge of an expert or on a professional decision.

Another method is the probability sampling, which can be further divided into three sub-methods:

- simple random sampling - It is based on the rules of like-hood estimate and presumes a random sample selection. The particular sample is selected using random numbers for element identification, all possible selections are equally probable. It can be used when the population of the domain inspected is homogeneous concerning the characteristics which need to be sampled. The results of the test may not be representative, as selected samples can be from just a part of the domain;
- multi-layer random sampling - It presumes the separation of population to layers which do not logically overlay. This sampling technique has a high potential on mean estimation and variation up against an un-layered strategy regarding the same population;
- semi-random sampling - Applies random selections to the initially sampled element (localization, characteristics) and also applies selection rules for all other elements. In practice, semi-random sampling uses the sampling grid, where the position of a grid is randomly determined or it is determined by regular spacing intervals. Within these grids, the parameters are evaluated, spatial tendencies are estimated, and regularities are concluded. This method assures the practical and relatively easy coverage of an inspected area.

Also, regarding probability sampling, we need to clarify another rule. The domain covered by a set of spatial data can be considered as a continuous space. If the space is divided in lots, we should give distinct attention to the omission and commission of elements, which fall partially or entirely in another lot. As we speak of data gathered from reality, a series of parameters will influence the quality of data (data source, the abilities of the operator, etc.). It is the duty of the data producer to define the homogeneous lots for selling the product by assuring the quality of the data.

All elements which belong to the selected sampling units are inspected. Elements from the dataset are compared with the definition domain in accordance to the measure plans selected.

2.3 Existing standards for inspection by sampling

Existing standards for inspection by sampling have not been elaborated for this domain of spatial data. Nevertheless, a big part of these methods and rules can be assimilated in our domain when taking samples.

We have the following standards at our disposal:
- ISO 2859 - 1 inspection from a continuous series of lots (ISO 2859-1);
- ISO 2859 - 2 inspection for individual or isolated lots (ISO 2859-2);
- ISO 2859 - 3 lot-skip inspection (ISO 2859-3).

The conformance quality level for a dataset is specified as AQL (Acceptance Quality Limit) in the ISO 3534-2 series (ISO 3534-2). It represents the rate of nonconforming articles at which a lot is rejected. It is particularly the risk of the data producer.

ISO 2859-2 is based on product specification and it operates with LQ (Limiting Quality), which represents the rate of nonconforming articles at which a lot is still accepted. This is known as the consumer's risk. Specification limits for determining the conformity for each element should be specified when ISO 2859 is applied based on product specification.

Additional series of ISO 2859 allow for multi-layer sampling, based on AQL or LQ. In terrestrial measurements or cadastre, based on these elements, we can establish (at the level of competent authorities) other types of sampling methods, based on the standard deviation for evaluating the positional accuracy. The precision and the consistency of the measurements will be evaluated on the field.

2.4 Guidelines to sample collection

Sample collecting is based on data from the original product, it has no relation with the source data. Selection is randomly made. During the selection, we must take care of the samples to be taken proportionally from each objective, objective class, evenly distributed on the whole inspected territory. A probable classification for collecting samples:

- checking the geometry - the objective geometry is erroneous if any point describing it is incorrectly positioned, or the dimensions of the objective have a deviation following the inspection measurements, which exceeds the tolerances defined for dimensions or positioning. The objectives of geometrical inspection consist of checking of natural positioning: right-angled checking of buildings, etc. Any deviation regarding the positioning, dimensions or geometrical conditions has the meaning of an error;
- checking the existence of the objectives - the existence of the objective is important in order to determine reality. It is considered an error if the objective is completely or partially missing, or if it is added to the physical reality;
- inspecting the attribute type values - it is considered an error if the attribute of a selected object is erroneous, is missing or it has an excess of data. Attribute data regarding the real estate, the parcels, the components, the cadastral number or the name are considered as missing if they don't show up, are investigated or are erroneous. Any omission of this type raises the number of errors;
- inspecting the topological links - during the topological inspection, any selection will be tested regarding the topological links, the extensions of the element. It is considered as a missing topological element if the element exists in reality, but has no topological links, or the links are erroneous OR the link points to a non existing objective OR the topological link exists, but the objective has no correspondence in the database.

Based on these inspections, the errors can be ranked based on the frequency of their appearance. We can rank the errors on the inspected objects as it follows:

- geometry;
- existence of the objective;
- topology;
- attribute-type value errors.

In case of inspected objects, we can say that we have errors with different ranks (multiple errors).

When establishing the quality, we make use of all inspection data (even those which are out of the tolerance), because only in this way we can make a correct evaluation.

Naturally, areas where the tolerance is found to be exceeded, based on the inspections, the remapping of that area is requested.

So basically, by using the sampling technique, instead of a population of N elements, we study a sample of n elements. We select the data sample (suggestive pattern), which represents the mean square root error usually equals the mean square root error of the population.

The volume of the data sample depends on:
- the size of the population (the database);
- the degree of confidence of the decision;
- economical reasons.

2.5 Total error of estimation

When estimating a parameter, the total estimation represents the difference between the calculated value of the estimator and the real value of this parameter (Suba, 2010b).

The total estimation error is due to:
- the sampling error
- the measuring error
- rounding of values or a characteristic division
- ”preconceptions” of the estimator

A part of the total error of estimation is also due to the following parameters:
- inconsistency of the inspected characteristics
- the random nature of sampling
- the known and accepted characteristics of the sampling plan

The sampling plans are affected by two types of errors:
- sampling error – is due to the samples which do not represent the inspected population with the necessary precision,
- measurement error – is caused by the characteristic measurement error on a sample which does not represent exactly the real value of the population i=σ.

It is desirable that sampling errors associated with any sampling plan, as well as measurement errors associated with the analyzing of the characteristics, to be quantified and reduced to a minimum.

The standard deviation in this case will be:

$$\sigma = \sqrt{\sigma_s^2 + \sigma_m^2} \quad (1)$$

where: σs – standard deviation of sampling,
σm – standard deviation of measuring.

The analyzed (measurement) error is usually negligible in comparison with the sampling error, because:

$$\sigma_m \leq \frac{\sigma_s}{3} \tag{2}$$

In this way we have:

$$\sigma \leq \sqrt{\sigma_s^2 \times \left(1+\frac{1}{9}\right)} \leq 1{,}05\sigma_s \tag{3}$$

This is executed in the case of very high data volumes, and we can inspect the following elements: the frequency of the objectives, the attribute-type data values, or the topological elements.

In case of spatial data, the five elements of quality will be evaluated differently than in the case of industrial product lots.
Positioning precision is evaluated by sampling of the guided zone or by sampling by judgment.

Completeness can be evaluated together with the thematic accuracy, and with the semi-random sampling of the products.
Logical consistency and time accuracy will usually be verified on the whole population.

2.6 Inspecting the area of the surfaces of the map

The purpose of inspection is to point out the calculus error of the area and redressing surface differences due to errors resulting from the real estate surface database, from map vectorisation, from actualisation of the maps or simply by evaluating the surface determination errors.

During the surface inspection, the area of the digital surface is compared against the area of the surfaces from the real estate inventory.

An inspection of the quality determination of the areas should be necessary. We proposed to make a presentation dealing with the errors of determining the areas through numerical calculus, having in mind the quality indicators of the field surveying (RMS error), or tolerances accepted by specialty forums.

We start from the analytic calculation method of the surface area:

$$S = 0.5\sum_{1}^{n}(X_{i+1} - X_{i-1})Y_i \tag{4}$$

or

$$S = 0.5\sum_{1}^{n}(Y_{i+1} - Y_{i-1})X_i \tag{5}$$

where: n – the number of jog points

The formulas presume knowing the coordinates of the jog points of the area to be calculated.

To evaluate precision, we will use the mean deviations:
- σs - RMS error of the area,
- σx, σy - RMS error of the coordinates.

Thus, from the RMS errors of the coordinates we can deduct the RMS error of the surface area, using the well known mathematical procedure:

$$\sigma_s = \left[\left(\frac{\partial S}{\partial X}\right)^2 \sigma_x^2 + \left(\frac{\partial S}{\partial Y}\right)^2 \sigma_y^2\right]^{\frac{1}{2}} \tag{6}$$

Having in mind the function of the analytical calculus of the surface area, we can calculate the differentials of the function with the following expressions:

$$\frac{\partial S}{\partial X} = 0.5\sum_{i=1}^{n}(Y_{i+1} - Y_{i-1}) = 0.5\sum_{i=1}^{n}\Delta Y_{i+1,i-1} \tag{7}$$

and

$$\frac{\partial S}{\partial Y} = -0.5\sum_{i=1}^{n}(X_{i+1} - X_{i-1}) = -0.5\sum_{i=1}^{n}\Delta X_{i+1,i-1} \tag{8}$$

Assuming that the mean errors of the coordinates are equal:

$$\sigma_x = \sigma_y = \sigma_c \tag{9}$$

we obtain the following relationship for the mean error of the surface area:

$$\sigma_s = 0.5\sigma_c\left(\sum_{i=1}^{n}\Delta X_{i+1,i-1}^2 + \Delta Y_{i+1,i-1}^2\right)^{\frac{1}{2}} \tag{10}$$

We can observe that:

$$\Delta X^2_{i+1,i-1} + \Delta Y^2_{i+1,i-1} = D^2_{i+1,i-1} \quad (11)$$

The mean error of the surface area will consequently be:

$$\sigma_s = 0.5\left(\sum_{i=1}^{n} D^2_{i+1,i-1}\sigma^2_{ic}\right)^{\frac{1}{2}} \quad (12)$$

where: σ_s - mean error of the surface area

σ_{ic} - mean error of determining the coordinates of point *i*

$D_{i+1,i-1}$ - the „cord" belonging to point P_i

n - number of jog points

We can observe that the mean error of the surface area depends on the precision of determining the jog points' coordinates. Further analyzing the expression, we can observe that the error of the surface area does not depend on the size of the surface, but it is related to the error of the jog points and the distance between points n+1 and n-1.

3. RESULTS

There are two key points for enhancing data quality in mapping applications – prevention and correction. Error prevention is considered being highly superior to error detection, because error detection is usually costly and does not guarantee a 100% success rate. The best way to prevent errors is to follow the regulations regarding the specifications, structure, dataset, implementation and evaluation procedures.

Thus, it can be concluded that setting out standards for mapping data contributors can prevent the appearance of the majority of errors contained in maps. Also, implementing the right sampling and error detection methods, the quality and reliability of existing maps can be evaluated and further enhanced.

In our test, we pointed out that, in order to reach a certain level of confidence regarding the data, it is not necessary to address the problem itself (in this case, to induce stricter standards or values for the representation of an area), but rather to address the source of the problem (the more precise determination of the joint points), which will solve the initial problem, and also contribute to the overall quality of the whole dataset.

4. DISCUSSIONS

In order to give an answer to the question contained in the introduction chapter - yes, it is possible to enhance the precision, the completeness and the temporal accuracy of the maps, but only with a moderate to great effort.

One problem that needs to be addressed is that usually users can contribute with data to existing datasets, but they cannot intervene on data that was provided by other users. This way, the whole error detection procedure (executed with the recommended sampling methods) is in the hand of only a few designated users, and after a marked error, the acceptance or rejection of the data is solely in their hands (Nistor, 2014). But this leads to other problems, as sampling errors will make us usually think a lot. The risk is due to the fact that instead of testing the dataset, we only test a part of it. According to a part of mathematicians, "the estimation error is the tribute paid for giving up on the complete inspection".

The standardization of the quality evaluation still has beneficial effects upon the users of the spatial data. User complaints are reduced, confidence in these data raises, rises and costs due to decisions based on erroneous decisions are reduced. A dataset can never be without errors. In order to obtain a high quality level, time, money and effort is needed.

The value of costs in an application with erroneous data can be estimated. (Example: during urban cadastral work, thematic accuracy and completeness have been evaluated erroneously. During an intervention, the pipe lines are represented erroneously or are not represented at all. The damage caused by the breaking of a heating pipe or a gas pipe can be calculated.). Part of this sum should be invested in quality management. The effect can be explained by raising the conformity level of the quality. Quality evaluation should not be a burden for the authorities, producers and data users. Instead, if possible, they should all contribute in order to obtain a high quality dataset.

5. REFERENCES

ISO 2859-1. Sampling procedures for inspection by attributes Part 1: Sampling schemes indexed by acceptance quality limit (AQL) for lot by lot inspection.

ISO 2859-2. Sampling procedures for inspection by attributes. Part 2: Sampling plans indexed by limiting quality (LQ) for isolated lot inspection.

ISO 2859-3. Sampling procedures for inspection by attributes. Part 3: Skip-lot sampling procedures.

ISO 3534-2 Statistics, Vocabulary and symbols. Part 2: Applied statistics.

ISO 3951-1 Sampling procedures for inspection by variables. Part 1: Specification for single sampling plans indexed by acceptance quality limit (AQL) for lot-by-lot inspection for a single quality characteristic and a single AQL.

Jakobsson, A. Giversen, J., 2009. Guidelines for Implementing the ISO 19100 Geographic Information Quality Standards in National Mapping and Cadastral Agencies. http://www.eurogeographics.org/documents/Guidelines_ISO19100_Quality.pdf (view at 28 feb. 2015).

Joos, G., 2006. Data quality standards. XXIII FIG Congress, Munich, Germany, october 8-13, 2006. https://www.Figure net/pub/fig2006/papers/ws02/ws02_03_joos_0906.pdf (view at 24 feb. 2015).

Nistor, S., Buda, A.S., 2014. Robust alternativ concerning the arithmetic mean and dispersion. *Journal of applied engineering sciences*, vol. 4(17), pp. 59-66, Oradea, Romania. http://www.arhiconoradea.ro/jaes/Journal_Archives/Revista_2014/JAES_MAI_2014/JAES_issue_1_mai2014.pdf.

Oteros, J. et al, 2013. Quality control in bio-monitoring networks, Spanish Aerobiology Network. *Elsevier - Science of the Total Environment*, 443(2013), pp. 559-565. http://www.uco.es/rea/publicaciones/andalucia/cordoba/J-Oteros_quality_13.pdf.

Suba, N.Sz., Suba, Șt., 2014. CAD and GIS interoperability - myth or possibility?. *RevCAD Journal of Geodesy and Cadastre*, vol. 17, pp. 133-136, Alba Iulia, Romania. http://www.uab.ro/geocad/upload/36_444_Paper17_RevCAD17_2014.pdf.

Suba, Șt., Suba, N.Sz., 2010a. Techniques of inspecting the quality of maps. *Analele Universității din Oradea – Fascicula de Construcții și Instalații Hidroedilitare (Annals of the University of Oradea - Bulletin of Constructions and Hidroedilitary Installations)*,pp. 253-260. http://www.arhiconoradea.ro/JAES/Journal_Archives/Istoric/Abstracts_Volum%2013_iulie2010.rtf.pdf.

Suba, Șt., Suba, N.Sz., 2010b. Data Quality Inspection by Sampling. *Analele Universității din Oradea – Fascicula de Construcții și Instalații Hidroedilitare (Annals of the University of Oradea - Bulletin of Constructions and Hidroedilitary Installations)*,pp. 279-288. http://www.arhiconoradea.ro/JAES/Journal_Archives/Istoric/vol.nov_2010supl.II.pdf.

van Oosterom, P., et. al., 2006. The core cadastral domain model. *Elsevier - Computers, Environment and Urban Systems*, vol. 30(5), pp. 627-660. http://dx.doi.org/10.1016/j.compenvurbsys.2005.12.002.

Volkan, C., Stubjaer, E., 2011. Design Research for Cadastral Systems. *Elsevier - Computers, Environment and Urban Systems*, vol. 35(1), pp. 77-87. http://www.yildiz.edu.tr/~volkan/publications/Design%20research%20for%20cadastral%20systems.pdf.

Zahir, A. et al, 2012. An integrated approach for updating cadastral maps in Pakistan using satellite remote sensing data. *Elsevier - International Journal of Applied Earth Observation and Geoinformation*, vol. 14(1), pp. 386-398. http://dx.doi.org/10.1016/j.jag.2012.03.008.