ON THE ROLE OF QUALITY RELATIONS IN PATCH-BASED SPATIAL DATA UPDATING

Arie CROITORU and Yerahmiel DOYTSHER
Technion – Israel Institute of Technology
Department of Civil Engineering, Division of Geodetic Engineering
Haifa 32000, Israel
ariec@tx.technion.ac.il , doytsher@geodesy.technion.ac.il

KEY WORDS: Updating, Accuracy, Collocation, Transformation, Distortion field model.

ABSTRACT

As spatial data becomes a central component in a variety of applications, the demand for up-to-date data is on the rise. In order to shorten the updating cycle time local updating is preferred, in which patches of up-to-date data must be incorporated into the existing data set. Although this can be done by using a global transformation model or a rubber-sheeting scheme, it is argued that in the case of patch-based updating the accuracy relations and its spatial variations must be considered. This requires adopting a field model for the various distortions in the data sets, as well as the implementation of proper computational tools. It is suggested that collocation can not only be used as such a tool, but that it may also encompass additional advantages, such as the ability to estimate the distortions in one data set based on the distortions in another data set. An example to the field model and the estimation of unknown signals in a data set are also presented.

1. INTRODUCTION

The demand for up-to-date spatial data has long been self-evident. Nowadays, As up-to-date spatial data are becoming a fundamental component in a variety of engineering, analysis and management operations and as this data is becoming readily available to a growing community of users, the requirement for up-to-date data is on the rise. In order to facilitate this requirement an updating process is employed. This process can be carried out either on a *global scale*, where the entire data set is replaced by a new up-to-date data, or on a *local scale*, where distinct areas in the existing data set are updated. Due to various drawbacks of the global scale updating process, such as its long duration and the considerable resources required for its implementation, a local updating process, during which only *patches* in the existing data are updated, is frequently preferred.

Both the global and the local updating schemes share several fundamental processing steps that are required for their success. These steps usually include *extracting up-to-date data*, *detecting and classifying changes*, and finally, *incorporating* the up-to-date data with the existing data. Up-to-date data extraction usually consists of processing up-to-date aerial photographs or a re-mapping of the interest area using a variety of surveying methods. Change detection and classification, which is at the heart of the updating process, consist of a comparison between the up-to-date and the existing data, resulting in a designation of areas or objects that were affected by change. The final step of incorporating the up-to-date data with the existing data usually consists of transforming the up-to-date into the existing data set using a variety of transformations. Most of the research effort in recent years was dedicated to the first two steps of the updating process, yet little attention was given so far to the problem of incorporating the existing data set and the up-to-date data (in the case of a local updating process, this includes incorporating several patches of up-to-date data).

The data incorporation problem is commonly solved by employing various geometric transformations. These transformations are realized by mathematical models with various degrees of freedom, ranging from a rigid-body transformation with three degrees of freedom up to an affine transformation with six degrees of freedom, or a projective transformation with eight degrees of freedom (Fagan and Soehngen, 1987). The transformation process begins with the measurement of homologous points in both data sets. If redundant points were identified the transformation parameters may then be estimated using the well known least squares adjustment technique, during which weights may be assigned to each measurement (Greenfeld, 1997[a]) ; (Greenfeld, 1997[b]). In the case of control points weights may be assigned by the rank of each point in the control network hierarchy (Greenfeld,

1997[b]). In case of a non uniform homologous point distribution a modified least squares scheme is required in order to eliminate the effect of leverage points (Kampmann, 1996). Additionally, various constraints may also be incorporated in order to maintain the consistency of the existing data.

Although a geometric transformation may bring both data sets into the same datum (thus eliminating the systematic effect), discrepancies between the overlapping area of the patch and the existing data are still likely to occur. This type of difficulty is also encountered during the vectorization process of scanned map sheets (Doytsher and Gelbman, 1995). In this process each map sheet is treated separately by vectorizing the required data in the map followed by a transformation (usually an affine transformation) of the resulting vector data using the state-plane coordinate grid that was overlaid in the map sheet. When several map sheets with overlapping boundaries are aggregated, discrepancies in the overlapping area still exist. This is caused by the inability of the affine transformation to account for the *random* part of the discrepancies between the two data sets. Although proper averaging of the overlapping vector data may eliminate the discrepancies, it may also introduce distortions in the vector data and by that violate the relationships between data elements (fixed length or angle, parallelism, perpendicularity, etc.) (Doytsher and Gelbman, 1995). Consequently, in order to account for the resulting random distortions a *rubber-sheeting* process is employed. During this process the distortions are spread linearly toward the center of the map sheet, where linearity is assumed along the boundary of the map as well as perpendicular to it (Doytsher, 2000). Doytsher (2000) also suggested a rubber-sheeting algorithm for non-rectangular map regions.

For patch-based updating purposes similar techniques may be employed. Although it may be assumed that both data sets share the same datum discrepancies are still likely to be found due to various factors such as the use of different datum points and the employment of different surveying techniques. These discrepancies, which contain systematic and random parts, can be eliminated using a proper transformation that accounts for the systematic part, followed by a rubber-sheeting that accounts for the random part of the discrepancies. Yet, by doing so several concerns should be noted:

- In the case of map sheets the purpose of the transformation is to bring each map sheet to the state-plane coordinate system. This is done by using the overlaid grid intersections in each map while assuming that the position of each intersection is errorless. Thus, each map is transformed separately based on pre-defined control points. This is not the case for an up-to-date data patch, where the purpose of the transformation is now to resolve the systematic part of the discrepancies between the information about the *same* datum as it is manifested in each data set. As this is done with homologous points from both sets it can not be assumed that one data set is errorless, and proper weight should be assigned according to the accuracy of each data set.

- For map sheets, the discrepancies that remain after the transformation are resolved by averaging. As it may be assumed that there is no extensive change in the accuracy of the data for the same map series, and therefore averaging the position of homologous points after the elimination of the systematic part of the discrepancies is permissible (in case of different map scales proper weights may be assigned). It may be argued that the same practice can be employed for a patch of up-to-date data that is to be incorporated into the existing data set. After proper transformation discrepancies are still likely to occur and they may be resolved by applying averaging and a rubber-sheeting process. Yet, averaging (and weight assignment) is not straightforward in this case since it is not clear whether both data sets share the same accuracy characteristics.

- During the rubber-sheeting process the distortions are spread linearly, as was described earlier. This can be justified by the assumption that the change in the distortions is linear throughout the data set. Yet due to surveying practices, map compilation techniques, and map sheet handling this may not be the case. Consequently, the distortions may not be uniform and may vary throughout the patch.

These concerns raise two fundamental questions that should be addressed prior to the incorporation of the up-to-date information, namely the *accuracy relations* between the data sets and the *spatial variation* of accuracy throughout the data set. In order to resolve these questions the sources as well as the behavior of the errors in a data set should be described. These will be discussed in the following section.

2. ERROR SOURCES AND THEIR SPATIAL "BEHAVIOR"

The sources of spatial data can vary from field surveying and photogrammetric techniques to digitizing existing map sheets. Following the work of Thapa and Burtch (1990), one may classify these sources into two different categories, namely *primary sources* and *secondary sources*. This classification is based on a differentiation between collection methods that use raw field measurement data, such as various surveying and photogrammetric techniques, and collection methods that rely on compiled measurements, such as digitized or scanned map sheets.

Each of these data sources can be characterized by a set of factors that contribute to the final accuracy of the data set. Primary methods are characterized by three major types of errors: personal errors; instrumental errors; and environmental errors, whereas secondary methods can be characterized by various error sources such as scanning or digitizing errors; compilation error; generalization error; and map material deformation errors (Thapa and Bossler, 1992). Additional detailed description of the error sources may be found in (Hunter and Beard, 1992); (Maffini et al. 1989). It should be noted that the error sources are associated with the accuracy of the data set as well as its quality, but while the accuracy is a statistical characteristic of the data set the quality of the data set depend on the context of use (Hunter and Beard, 1992). Additionally, the term accuracy and error can be related to several aspects of the data. Usually these aspects are classified into positional, attributive, topological, and temporal aspects (Shi, 1998). In the scope of this paper the terms accuracy and error are related only to the positional accuracy of the data.

As the accuracy of the spatial data and its products is of great importance to its providers as well as to its end users, considerable research effort was dedicated to the study of errors and error propagation in spatial data. In the context of vector data, a substantial part of this effort was dedicated to the modeling of errors for vector data entities such as points and lines (for example the work of Easa (1995); Stanfel and Stanfel (1993); Stanfel (1999)) and to the modeling of errors of spatial operators such as overlay operations (for example the work of Chrisman (1989); Veregin (1990); Leung and Yan (1998)).

Although the error sources can be identified and even modeled in some cases, usually the magnitude and particular behavior of each error source cannot be specified for an existing data set (Ehlschlaeger and Goodchild, 1994). Even if accuracy information exists for a specific data set, it is usually vague or given in general terms. Consequently, the errors must be assessed in other ways, which are usually empirical or designed for a specific context (Ragia and Winter, 2000). This is usually done by utilizing a *reference data set* in the form of a ground-truth data set or another data set. Once such a reference is obtained, deviations of the examined data set from the reference data set can be found, and its validity can be verified using various statistical tools, such as the ANOVA tool. If several data sets covering the same area are available, the deviation values between data elements may also serve as an assessment tool (Ragia and Winter, 2000).

The error models described fall short of providing a sufficient solution to the problem of spatial error modeling for two primary reasons. The first is related to the inability of these models to support complex spatial objects such as topographic databases (Goodchild et al. 1992). The second reason is related to the inability of the models to take into account the spatial autocorrelation of the various error sources (Ehlschlaeger and Goodchild, 1994). This led to the development of a more generic error model in the form of an *error field* or a *distortion field*, which is a continuous description of the spatial variation of the error. Examples to such models may be found in the work of Goodchild at al. (1992), Ehlschlaeger and Goodchild (1994), Hunter and Goodchild (1996), and Church et al. (1998).

Using a field model, the errors or distortions may also be described as *signals*. The signals are generated by the various error sources, where each source may be characterized by its own magnitude and spatial behavior. The final signal at each point in the field is the sum of all signals present, thus in a given point the signals represent an "overall" error. An example to such signals may be seen when comparing a vector data set of roads with an aerial image covering the same area. The signal sources in this case are the exterior orientation parameters of the image, the projective geometry of the camera, and the topography of the area (relief displacement).

An example that illustrates this is given in figure 1. To demonstrate the nature of the signals a hilly build-up area was selected. An aerial image of the area (Fig. 1a) was rectified to fit a vector data set of roads (Fig. 1b) by a four-parameter Helmert transformation (two translations, a rotation and an overall scale factor). The transformation parameters were estimated using a set of eight conjugate points (marked by the crosses in Fig. 1a). This was followed by a manual digitizing of the location differences (residuals) between the roads in the

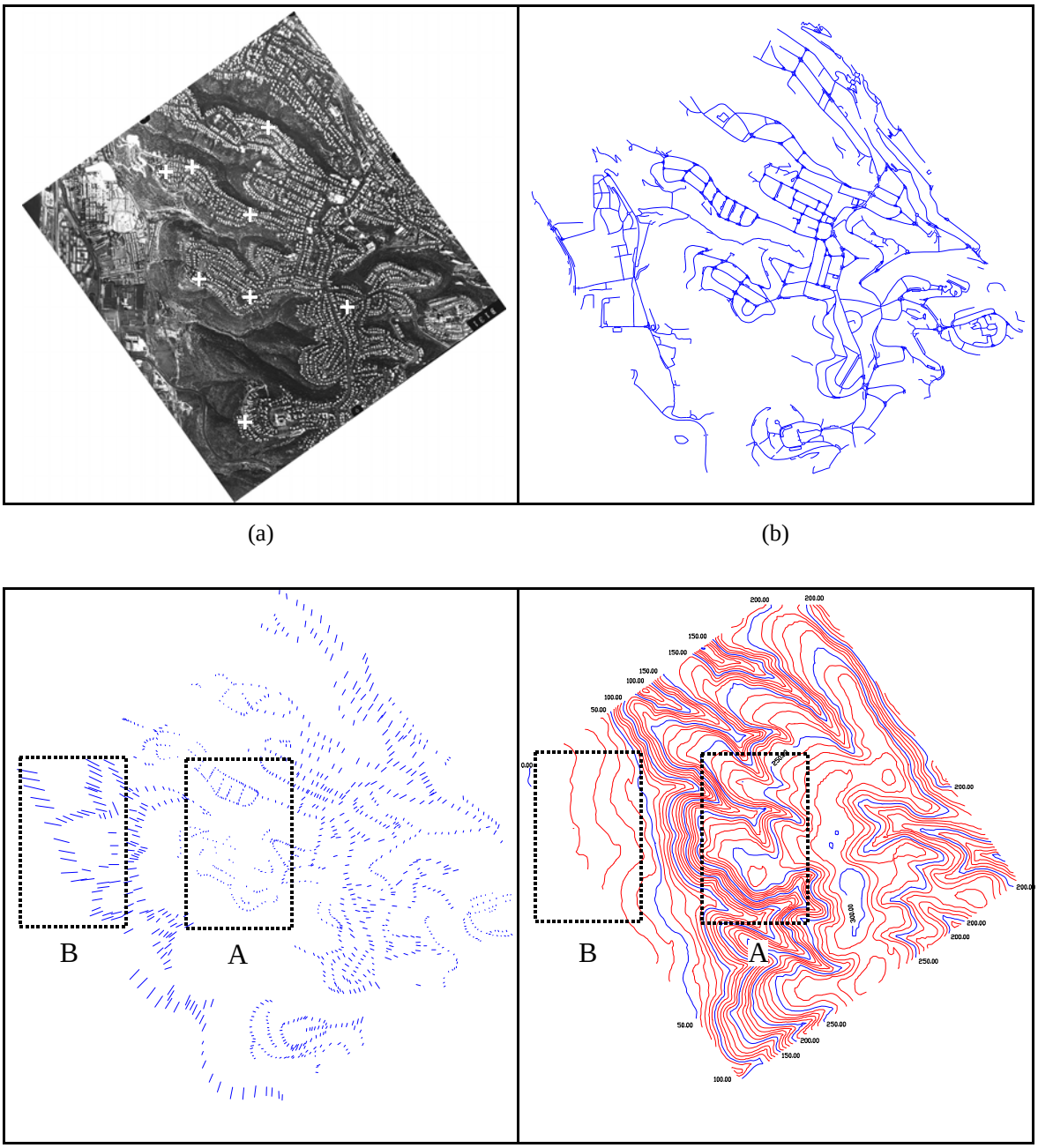

Figure 1: An example to signals in an aerial image – (a) the original image (the crosses mark the location of the points used for rectifying the image); (b) The corresponding data form a roads layer; (c) The signal that was measured; (d) a DTM covering the aerial image area.

vector data and the rectified image (Fig 1c). Since the points used for the rectification were on high grounds, the image was rectified to fit these areas. A comparison between the magnitude and direction of the signal and the topography of the area shows this – while the signal on high grounds is relatively low (area A in Fig. 1c, Fig. 1d), it becomes higher when moving to lower grounds (area B in Fig. 1c, Fig. 1d). It is also important to note that near signals are correlated to each other as can be seen from their magnitude and direction, but for distant signals this correlation is low. Thus, the spatial correlation of the signals is apparent.

3. INCORPORATING THE SIGNALS INTO THE TRANSFORMATION MODEL

One of the means to take into consideration the existence and the statistical nature of signals is the employment of the *Collocation* technique. Given a set of observations (l) and a parametric model (A) with unknowns (x), in this technique the residuals obtained from the classical least squares model are decomposed into a set of *signal* components (s) and a set of *random* components (n) (Moritz, 1972):

$$l = Ax + s + n\,. \tag{1}$$

A least-squares solution to eq. (1) is given by (Moritz, 1972) ; (Cross, 1983):

$$\begin{aligned} \hat{x} &= \left(A^T\left(C_s + C_n\right)^{-1}A\right)^{-1}A^T\left(C_s + C_n\right)^{-1}l \\ \hat{s} &= C_s\left(C_s + C_n\right)^{-1}\left(l - A\hat{x}\right) \\ \hat{n} &= C_n\left(C_s + C_n\right)^{-1}\left(l - A\hat{x}\right) \end{aligned}\,, \tag{2}$$

Where C_s and C_n are the variance-covariance matrices for the signal and the noise respectively.

Implementing collocation requires knowledge of the parametric model as well as the variance-covariance matrices. The parametric model is usually in the form of a geometric transformation, wherein the degrees of freedom are chosen according to prior knowledge. if such knowledge is not available a Helmert transformation is always permissible as a starting point (Buiten 1978). The covariance matrices can be estimated in some cases based on the properties of the problem at hand, but in most cases they are estimated empirically from the data at hand (Mikhail, 1976). A model for describing the covariance function *C* between two points, which is frequently used for transformation problems as well as photogrammetric applications is given by the Gaussian model (Mikhail, 1976):

$$C = k_1 e^{-k_2 d^2}\,, \tag{3}$$

Where $\boldsymbol{d}$ is the distance between the points, and $\boldsymbol{k_1}$,$\boldsymbol{k_2}$ are constants. It should be emphasized that the covariance matrices, which describe the spatial dependency of the signal or noise values between two points, are at the core of the collocation technique. As these matrices describe the statistical "behavior" of the signals and the noise, they provide the means of incorporating this information into the adjustment model (Deakin et al., 1994).

In view of the concerns discussed in sections 1 and 2, an analysis of the collocation technique in the context of patch-based updating indicates a few potential advantages over other transformation techniques:

- The statistical nature and the spatial dependency of the distortions is accounted for in the collocation solution via the covariance matrix. Hence, linearity is not assumed and the value of the signal (distortion) at each point can be estimated.
- Using the collocation technique, it is possible to estimate the value of the signal where no conjugate points exist. This is an advantage for an updating process since it enables estimating the signal in points that does not exist in the current data set.
- When transforming data set *A* with signals $\boldsymbol{s^B}$ to data set *B* with signals $\boldsymbol{s^A}$ the resulting signal $\hat{S}^T$ estimated by the collocation is the sum of the signals in each data set (Buiten, 1978):

$$\hat{S}^T = s^A + s^B\,. \tag{4}$$

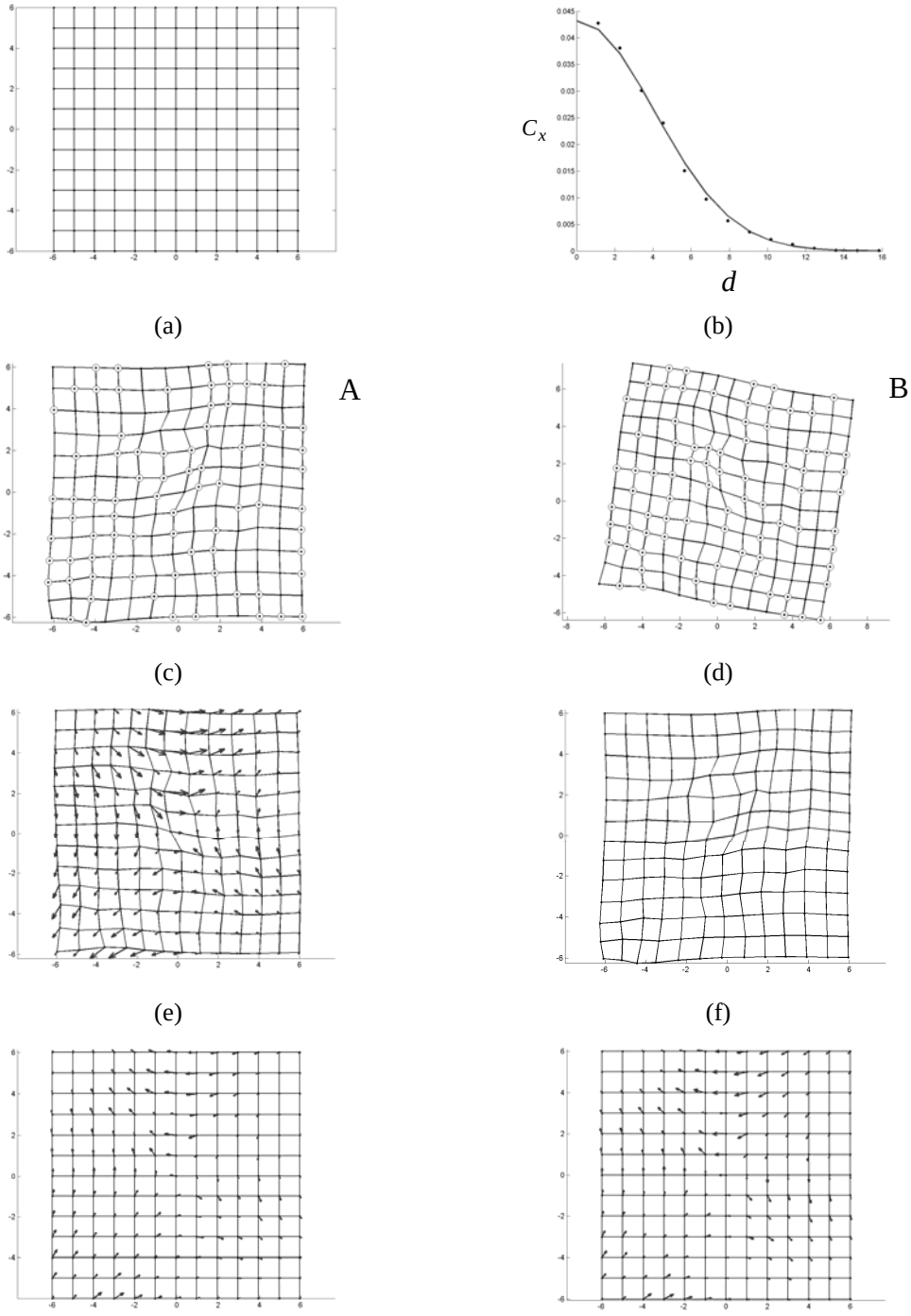

Figure 2: Recovering an unknown signal of a data set by collocation: (a) the original grid; (b) the covariance function for the signal in x direction; (c) grid A (circles mark points used for the computation); (d) grid B (circles mark points used for the computation); (e) grid B transformed to grid A with an estimation of the total signal (marked by arrows); (f) the transformed grid B after applying the total signal; (g) the recovered signal for grid B; (g) the original signal that was applied to grid B.

Although the last potential advantage can be considered a shortcoming of collocation in some cases, it could hold an important advantage in the case of updating. In view of the fact that for an existing data set the signals $\boldsymbol{s}^{\boldsymbol{B}}$ are usually unknown, this provides the means to estimate $\boldsymbol{s}^{\boldsymbol{B}}$ provided the signals $\boldsymbol{s}^{\boldsymbol{A}}$ in the up-to date data can be estimated (this may be assumed as the up-to-date data is the product of a recent data acquisition and processing). If both signals are recovered, they can be then used to estimate the positional accuracy of each data set.

An example of this last capability can be seen in Fig. 2. In a simulation, a grid of 13 by 13 points (Fig. 2a) was distorted by an artificial two-dimensional signal function. An example of the covariance function for the signal in $\boldsymbol{x}$ direction can be seen in Fig. 2b (estimated by the model described in eq. 3). The signal was applied twice to the grid with different parameters, resulting in two different distorted grids – grid A (Fig. 2c) and grid B, which was also transformed by a rigid body transformation (Fig. 2d). No random errors were introduced.

In this example it was assumed that the signal for grid A is known and an attempt to recover the signal for grid B was made. Using collocation, grid B was transformed to grid A using a rigid body transformation model. Only 80 of the grid points (the circled grid points in Fig. 2c and Fig 2d) were used for the computation of the transformation parameters and the signals. The transformed grid B and the estimated total signal (marked at each grid point) can be seen in Fig. 2e. The total signal was then applied to grid B resulting in a grid similar to grid A (Fig. 2f, for comparison see Fig. 2c). As it was assumed that $\boldsymbol{s}^{\boldsymbol{A}}$ is known, it was possible to recover $\boldsymbol{s}^{\boldsymbol{B}}$ (Fig. 2g). For a qualitative assessment of the results, this can be compared to the original signal that was applied to grid B (Fig. 2h).

4. CONCLUDING REMARKS

The concepts described in this paper are the basis of an on-going research toward the formulation of a tool aimed to facilitate an optimal fusion of the existing and the up-to-date spatial data sets. As the basis for such a fusion is the accuracy relations between data sets, such a tool must provide the ability to establish these relations even if there is little information regarding the accuracy of one of the data sources. Furthermore, as spatial data may not have homogenous accuracy this tool should be able to take into account the spatial variations of the accuracy using a proper error model.

For this purpose, an attempt to analyze collocation as a possible tool was made. It was shown that if the signal on one data set is known, it is possible to estimate the signal in another data set, for which the signal is unknown, thus providing the means to estimate the quality relations as well as to account for the spatial variation of the error. Implementing this tool for updating purposes requires future work, mainly in the development of proper estimation scheme for covariance matrices, and in the incorporation of such a tool into a patch-based updating process.

ACKNOWLEDGEMENTS

The authors would like to thank the Survey of Israel for providing the aerial images and the corresponding vector data sets that were used in this work.

REFERENCES

Buiten, H. J., 1978. “Junction of nets by collocation”. Manuscripta Geodetica, Vol. 3, pp. 253-297.

Chrisman, N. R., 1989. “Modeling error in overlaid categorical maps”. “Accuracy of spatial databases”, (Michael Goodchild and Sucharita Gopal, Eds.), Taylor & Francis, 290 pages, pp. 21-34.

Church, R., Curtin, K., Fohl, P., Funk, C., Goodchild, M. F., 1998. “Positional distortions in geographic data sets as a barrier to interoperation”. ACSM annual convention, pp. 377-387.

Cross, P. A., 1983. “Advanced least-squares applied to position fixing”. Working paper No. 6, North east London Polytechnic, Department of Land Surveying, 205 pages.

Deakin, R. E., Collier, P. A., Leahy F. J., 1994. “Transformations of coordinates using least squares collocation”. The Australian Surveyor, March 1994, pp. 6-20.

Doytsher, Y. 2000. “A rubber sheeting algorithm for non-rectangular maps”. Computers & Geosciences, 26 (2000), pp. 1001-1010.

Doytsher, Y., Gelbman, E., 1995. “Rubber sheeting algorithm for cadastral maps”. Journal of Surveying Engineering, November 1995, pp. 155-162.

Easa, S., 1995. “Estimating line segment reliability using Monte Carlo simulation”. Surveying and Land Information Systems, Vol. 55 (3), pp. 136-141.

Ehlschlaeger, C. R., Goodchild, M. F., 1994. “Uncertainty in spatial data: defining, visualizing, and managing data errors”. Proceedings of LIS/GIS ’94 annual conference, pp. 246-253.

Fagan, G. L., Soehngen, H. F., 1987. “Improvement of GBF/DIME file coordinates in a geobased information system by various transformation methods and “rubbersheeting” based triangulation”. Proceedings of Auto-Carto 8, eighth international symposium on computer-assisted cartography, pp. 481-491.

Goodchild, M. F., Guoqing, S., Shiren, Y., 1992. “Development and test of an error model for categorical data”. Int. J. Geographic Information Systems, Vol. 6 (2), pp. 87-104.

Greenfeld, J. 1997[a]. “Consistent property line analysis for land surveying and GIS/LIS”. Surveying and Land Information systems, Vol. 57 (2), pp. 69-78.

Greenfeld, J. 1997[b]. “Least squares weighted coordinate transformation formulas and their application”. Journal of Surveying Engineering, November 1997, pp. 147-161.

Hunter, G. J., Beard K., 1992. “Understanding error in spatial databases”. The Australian Surveyor, Vol. 37 (2), pp. 108-119.

Hunter, G. J., Goodchild, M. F., 1995. “Dealing with error in spatial databases: A smiple case study”. Photogrammetric Engineering & Remote sensing, Vol. 61 (5), pp. 529-537.

Hunter, G. J., Goodchild, M. F., 1996. “A new model for handling vector data uncertainty in geographic information systems”. URISA journal, Vol. 8 (1), pp. 51-57.

Kampmann, G., 1996. “New adjustment techniques for the determination of transformation parameters for cadastral and engineering purposes”. Geomatica, Vol. 50 (1), pp. 27-34.

Leung, Y., Yan, J., 1998. “A location error model for spatial features”. Int. J. Geographical Information Science, Vol. 12 (6), pp. 607-620.

Maffini, G., Arno, M., Bitterlich, W., 1989. “Observations and comments on generation and treatment of error in digital GIS data”. “Accuracy of spatial databases”, (Michael Goodchild and Sucharita Gopal, Eds.), Taylor & Francis, 290 pages, pp. 55-67.

Mikhail E. M., 1976. “Observations and least-squares (with contributions by F. Ackermann)”. IEP-Dun Donnolley, New-York, 497 pages.

Moritz, H., 1972. “Advanced least-squares methods”. Reports of the department of geodetic science No. 175, the Ohio State University, 129 pages.

Ragia, L., Winter, S., 2000. “Contributions to a quality description of aerial objects in spatial data sets”. ISPRS journal of Photogrammetry & Remote Sensing, Vol. 55 (2000), pp. 201-213.

Shi, W., 1998. “A generic statistical approach for modeling error in geometric features in GIS”. Int. J. Geographical Information Science, Vol. 12 (2), pp. 131-143.

Stafnel L. E., Stafnel, C. M., 1993. “A model for the reliability of line connecting uncertain points”. Surveying and Land Information Systems, Vol. 53 (1), pp. 49-52.

Stafnel, L. E., 1999. “Line segment reliability”. Proceeding of ACSM/WFPS/PLSO/LSAW, pp. 173-190.

Thapa, K., Bossler J., 1992. “Accuracy of spatial data used in geographic information systems”. Photogrammetric Engineering & Remote sensing, Vol. 58 (6), pp. 835-841.

Thapa, K., Burtch, R. C., 1990. “Issues of data collection in GIS/LIS”. Proceedings of ACSM/ASPRS annual meeting, Vol. 3, pp. 271-283.

Veregin, H., 1995. “Developing and testing of an error propagation model for GIS overlay operations”. Int. J. Geographical Information Science, Vol. 9 (6), pp. 595-619.