Integration of multiple collected polygons with a raster-based approach

V. Walter [1,*]

[1] Institute for Photogrammetry, Universität of Stuttgart, Germany – volker.walter@ifp.uni-stuttgart.de

Commission IV, WG IV/3

KEY WORDS: Data Integration, Conflation, Data Quality

ABSTRACT: In this paper, we present an approach for the integration of multiple collected (vector-)polygons that is computed in the raster domain. In a first step, all polygons are transferred to the raster domain with a vector/raster-conversion. The integration in the raster domain is a simple pixel-wise summation that is much simpler than comparable approaches in the vector domain. The results can be optimized with image processing operators. Finally, the integrated data are transferred back into the vector domain with a raster/vector-conversion. This approach can integrate not only 2 datasets but is also able to integrate *n* datasets without any modification. We will demonstrate this approach on data that was multiple collected in a student project and we will discuss how the integration results can be evaluated with quality measures.

1. INTRODUCTION

In the past decades, many algorithms were developed for the integration of spatial data from different sources. A very good overview is presented in (Xavier et al., 2016). Practically all existing algorithms work in the vector domain with the exception of (Seo and O’Hara, 2009) who presented a method for assessing the geometric quality of linear spatial vector data by converting them first in the raster domain and calculating displacement vectors based on raster buffers. This approach is very efficient but it is not able to integrate the data. To the best of our knowledge, all existing integration algorithms can conflate only two datasets at one time. The simultaneous integration of *n* datasets into one common datasets is so far not described in the literature.

An application that need to integrate not only two but more datasets could be for example the extraction of patterns from data from mobile sensors: Large numbers of pedestrians or cars equipped with mobile sensors can collect continuously spatial data. The result is a massive multiple collection of geospatial data. With an integration, we can extract for example movement patterns or even the geometry of streets or other spatial objects. For example, Sultan et al. 2017 describe an approach that uses GPS trajectories collected by multiple cyclists to extract spatial patterns.

Another application area is the collection of spatial data with paid crowdsourcing. It has been shown that we can collect high quality data with paid crowdsourcing (Walter and Sörgel 2018). However, the problem is that the results can be extremely inhomogeneous. Parts of the data are collected with high quality whereas other parts are collected with very low quality or even with senseless data (Bernstein et al. 2010). This kind of problem does not exist in the same intensity in Volunteered Geographic Information (Goodchild 2007) projects like OpenStreetMap (OSM), because there is a quality control of other users. If a user adds incorrect data, it is very likely that other users will recognize that and correct them after a while (this must not happen immediately but can also happen after some months or even years). Additionally, the completeness of OSM datasets increases over the time because more and more data are added until (most of) all objects are collected (Barron et al. 2013).

An approach to collect homogenous high quality data also with paid crowdsourcing would be to collect the data not only once but multiple times and to improve the quality with an integration of these datasets. This follows the idea of the Wisdom of Crowds. Charles Darwin’s cousin Francis Galton first observed the Wisdom of Crowds in 1907 (Shrier et al. 2016). He found out that the average guess of the weight of an ox in a ‘guess the weight of the ox’ competition was nearly accurate. The average judgement converges to an optimum result. In average, the participants estimated the weight of the ox was 1207 pounds. The real weight was 1198 pounds

The typical approach to integrate two spatial datasets in the vector domain is that first identical geometries are identified (matching) and then the actual integration takes place (Walter and Fritsch, 1999). Integration techniques of this kind are also known under the name conflation, which comes from the Latin *con flare* meaning "blow together" (Lynch and Saalfeld 1985). For this purpose, the perpendicular distances from all intermediate points from one geometry to the matched geometry and vice versa are calculated. The integrated geometry is then calculated by connecting all center points of all perpendicular distances (Volz and Walter 2006). That means that if we integrate two polygons P_1 and P_2, which have n_1 and n_2 intermediate points, the resulting integrated polygon will have $n_1 + n_2$ intermediate points.

If we want to integrate more than two datasets (see for example Figure 1) this approach has problems. If we would select one start dataset and then integrate successive all other datasets to this dataset, we would get different results depending on which dataset was selected as start dataset. In addition, an iterative approach where we first integrate dataset 1 with dataset 2 and then the integrated dataset with dataset 3 and so on, would lead to different results depending on the processing order. Furthermore, the integrated polygons would contain a huge

* Corresponding author

number of intermediate points and could contain self-overlaps. In order to solve this problem, we introduce a raster-based integration approach which is very simple to realize and which is able to integrate any number of datasets.

The rest of the paper is organized as follows. In section 2, we describe the implemented workflow and the corresponding parameters. In section 3, we show the test data and the reference data that we used for testing the approach. In the next section, we show integration results on examples and discuss the influence of the implemented parameters. In section 5, we discuss how the integration results can be evaluated with similarity measures. A conclusion and outlook completes the paper.

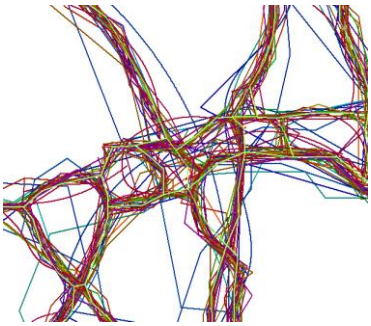


Figure 1. Example of multiple collected road network data

2. WORKFLOW

The workflow of the integration is shown in Figure 2. First, we transform the vector data into a raster cell matrix with a vector/raster-conversion. As parameter, the cell size (in meter) has to be defined. Initially, all raster cells have the value "0". The vectors are “drawn” onto the matrix and each time the corresponding cells are incremented by one (see Figure 2 a-c).

At those locations where the geometries of the different datasets are similar, we will get raster cells with high values – at all other locations low values. The raster cell matrix can then be processed with image processing methods, like smoothing, morphological operations or binarization (see Figure 2 d). In our approach, we first normalize the data and then we apply a 3x3 Gauss filter. The final computation in the raster domain is a binarization with a selectable threshold. The raster representation is then skeletonized and transformed back into a vector representation (see Figure 2 e). Finally, the vector data is smoothed with a Douglas-Peucker line smoothing (Douglas and Peuckcr 1973) which needs a buffer width as input parameter.

An advantage of this approach it, that is can be used for line and area objects in the same way since we describe the area objects by their surrounding lines in the vector and in the raster domain. Alternatively, it would be thinkable to represent the areas not only with their boundaries but also to use the interior pixels. The process chain could be adapted easily to this kind of representation.

3. DATASETS

Figure 3 shows a map representing the sixteen states of Germany which was used as a source for data collection. We advised 23 students to collect all states as part of a student exercise.

Figure 4a shows a part of the input map and Figure 4b shows the same part of the data that were collected by the different students. It can be seen that the different students collected the data very differently. Some of the students collected the data very precisely whereas other students collected the data only with very few intermediate points. Figure 4c shows the reference data set which we collected by ourselves

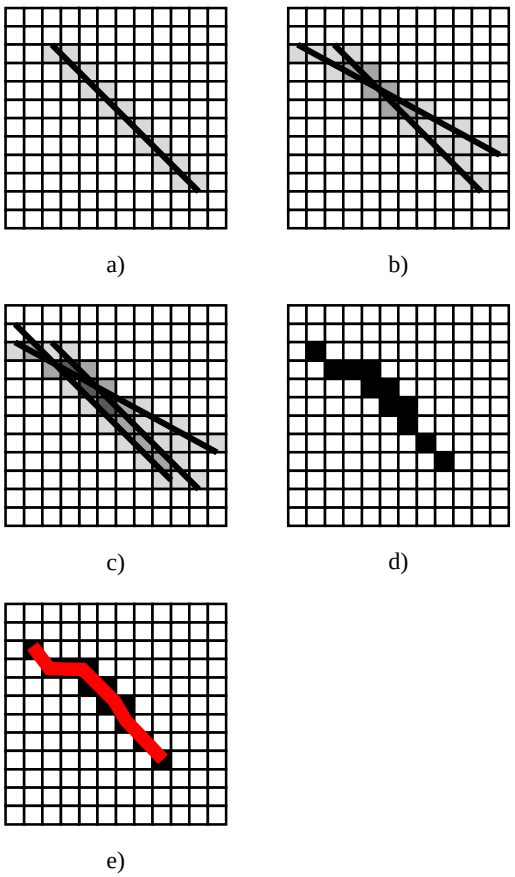
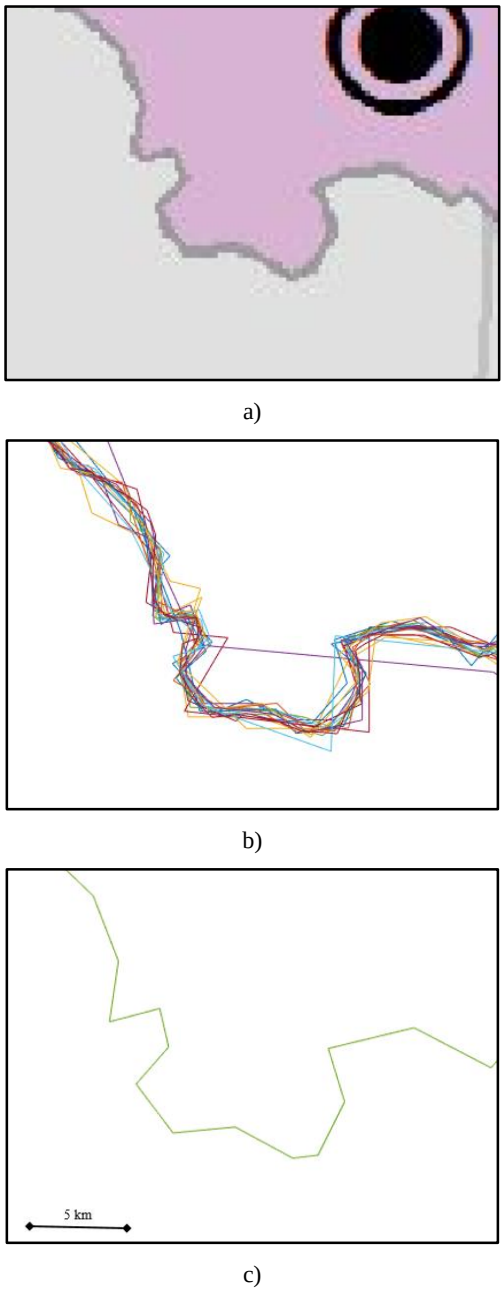


Figure 2. Workflow of the raster-based integration

Figure 3. Data source for the collection of the data



a)

b)

c)

Figure 4. Used datasets: (a) Zoom-in into the data source that was used for data collection (b) Results of multiple data collection (c) Reference data set

4. RESULTS

Figure 5 shows on examples the processing of the data in the raster domain. The map from which the data was collected is shown in Figure 5a, the reference data in Figure 5b and the multiple data collection in Figure 5c. Figure 5d, 5e and 5f show the corresponding raster data with different raster cell sizes: 300m, 150m and 75m. The images are normalized into an interval [0, 255]. Large cell sizes (like in Figure 5d) have a smoothing effect and the different polygons from the multiple data collection cannot be distinguished anymore in the raster domain. If we use smaller cell sizes, we can see the structure of the different polygons still in the raster data. This effect can be seen already in Figure 5e and even stronger in Figure 5f.

Therefore, the next processing step is a smoothing with a Gauss filter to eliminate the high frequency structures. We used in all examples a 3x3 filter mask. Figure 5g, 5j and 5h show the results of the smoothing of the data of Figure 5d, 5e and 5f. It can be seen that in Figure 5h (with raster cell size 75m) it is still possible to identify the different polygons of the input data. This is an indicator that a larger filter mask should be used for the smoothing.

After the smoothing, the data is binarized and the skeleton is computed. The results of this step can be seen in Figure 5j, 5k and 5l. The skeleton in Figure 5l is not as smooth as in the other examples because of the insufficient smoothing of the data in the step before.

After calculating the skeleton, the data has finally to be transferred back into the vector domain. This step can be seen in Figure 6. The map from which the data was collected is shown in Figure 6a. The selected part of the map is the same as in Figure 5 but a larger area can be seen. The extension of the data of Figure 5 is indicated with a rectangle. Figure 6b shows the multiple collected data and Figure 6c the result of the processing in the raster domain. For this example, we used a cell size of 100m and again a smoothing with a 3x3 Gauss filter.

Figure 6d shows the result of the raster/vector conversion in blue and the reference data in red. After a raster/vector-conversion, usually a line smoothing has to be applied. Otherwise, the data can contain high frequency structures that we do not want to have in the data. Examples can be seen in the upper left part of the test area. The line smoothing should also eliminate blocky structures because of the quadratic pixels. For the line smoothing, we use a Douglas-Peucker algorithm which recursively divides the lines into smaller parts and eliminates all points that are in a buffer with a width ε. In Figure 6e we used a buffer width of 1000m and in Figure 6f a buffer width of 2000m. The larger the buffer width the larger is the smoothing effect.

5. QUALITY EVALUATION

When we compare visually the results with the reference data, it can be seen that the integrated data show in many areas less details as the reference data. This is not because of the line smoothing with the Douglas-Peucker algorithm but is inherent to the proposed approach. It can be seen already in the input data that small details are “washed out” because of the multiple geometries. That means that the result of the integration of multiple collected data will be typically in a larger scale as the input data itself.

The quality of the results depends of course strongly on the selected parameters of the process chain, which are in our case: the raster cell size, the size of the Gauss filter, and the smoothing factor of the Douglas-Peucker algorithm. In order to optimize those parameters we calculated different integrated datasets with different parameters and compared the results with the reference data. We measured the similarity of the integrated data and the reference data with the following measures:

- Area difference
- Hausdorff difference
- Centroid point difference
- Perimeter difference

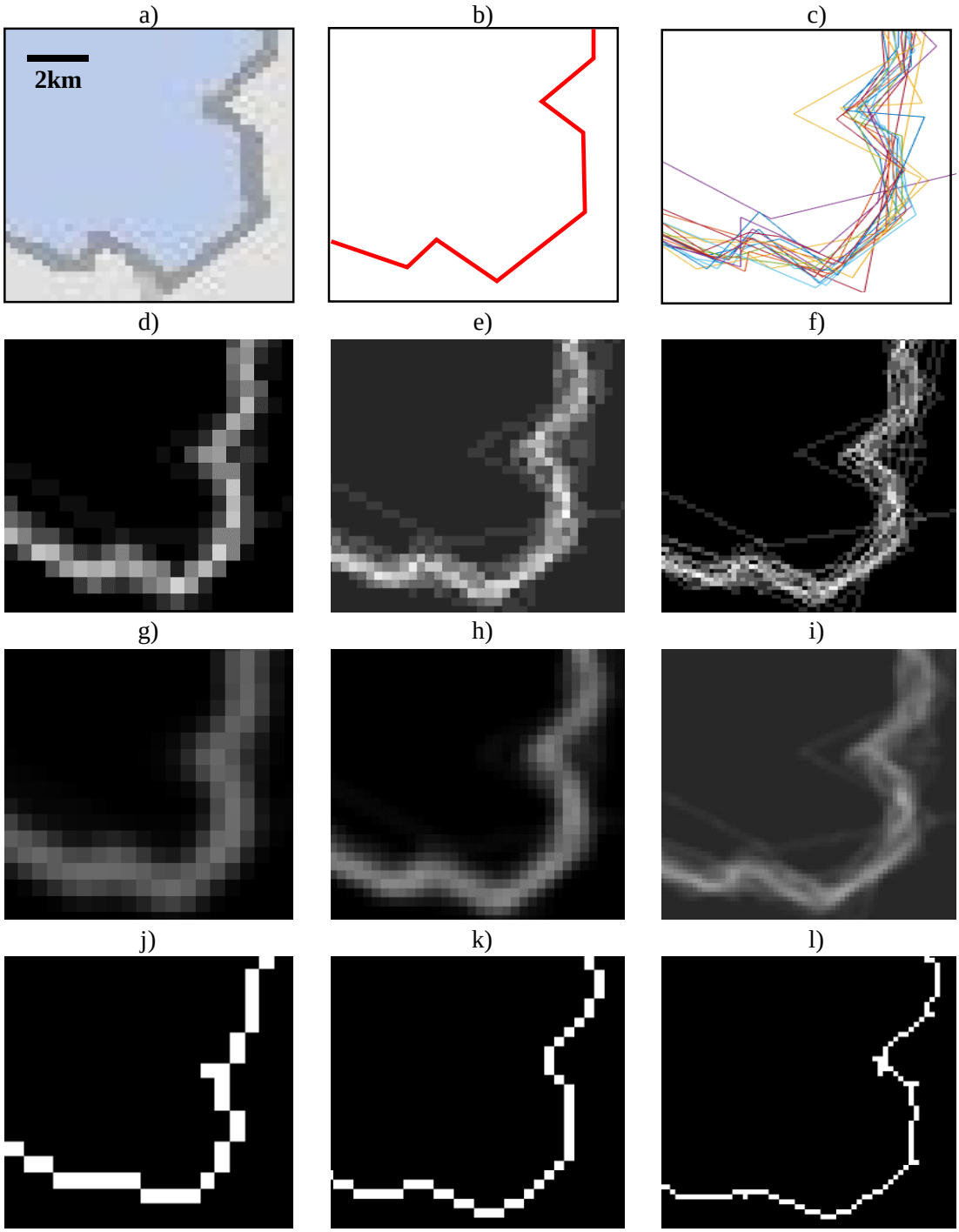


Figure 5. Results of the raster-based integration: (a) data source; (b) reference data; (c) multiple data collection; (d) – (f) raster data before smoothing with raster cell sizes 300m, 150m and 75m (g); raster data after smoothing with a 3x3 Gauss filter with raster cell sizes 300m, 150m and 75m; (j) – (l) skeleton with raster cell sizes 300m, 150m and 75m

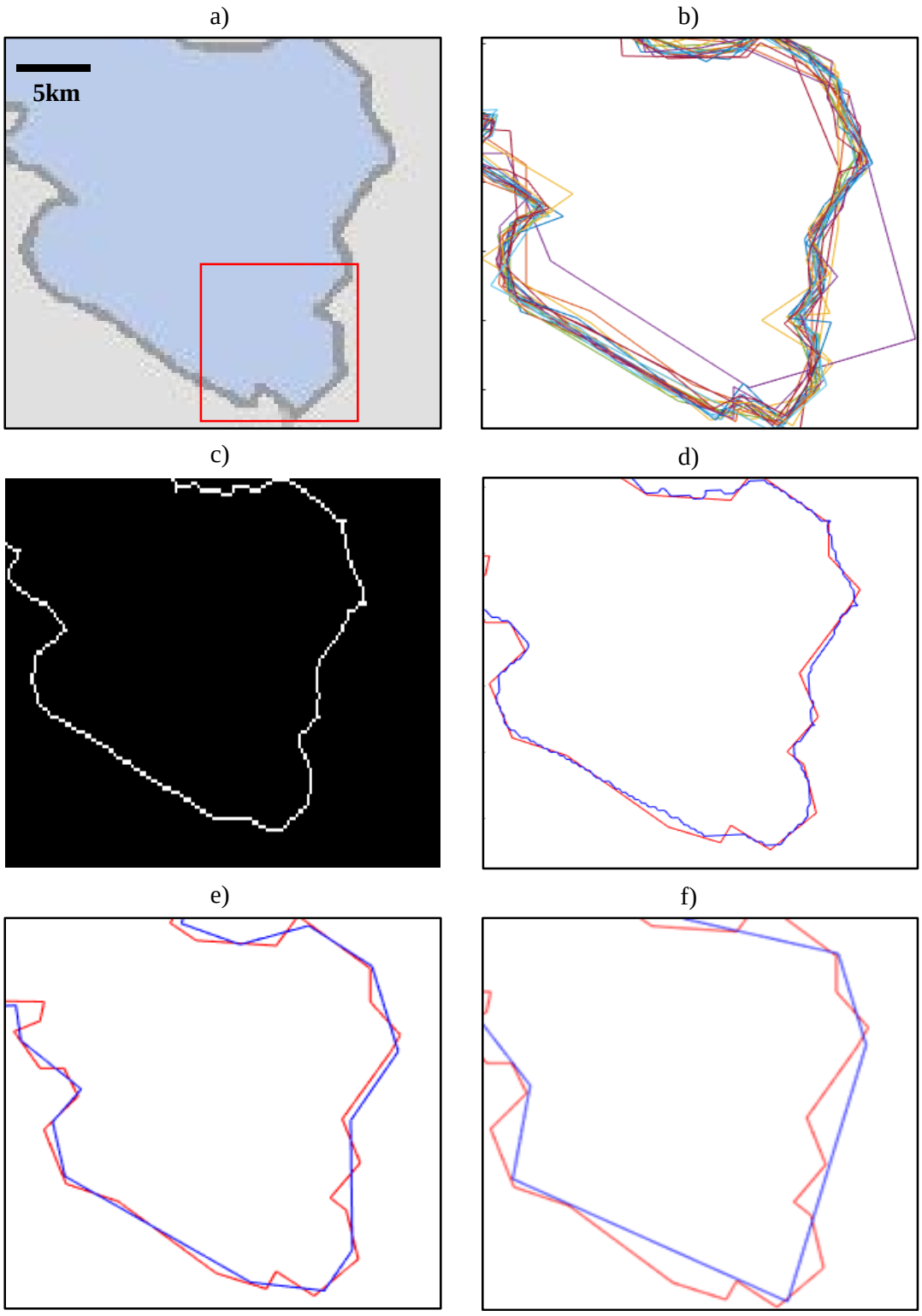


Figure 6. Results of the raster-based integration: (a) data source; (b) multiple data collection; (c) skeleton (d) result of the raster/vector-conversion in blue, reference data in red; (e) vector data smoothed with Douglas Peucker with buffer width 1000m in blue, reference data in red; (f) vector data smoothed with Douglas Peucker with buffer width 1000m in blue, reference data in red

In order to evaluate not each single polygon, we measured the average differences of all polygons of the whole dataset (16 polygons).

Table 1 shows exemplarily the results by using the cell sizes 150m, 110m and 75m. For the 75m raster cell size we used two different Gauss filters with a size of 3*3 and 9*9 because the 3*3 filter does not smooth the data enough for small raster cell sizes as we discussed already in the section above. In all configurations we applied no line smoothing. In the table we marked the highest similarity with green colour and the lowest similarity with red colour for each similarity measure. It can be seen that there is no clear trend detectable. If we change the parameters in the process chain, typically the similarity measures improved for some of the polygons but declined for other polygons at the same time.

cell width	**150m**	**110m**	**75m**	**75m**
filter size	**3x3**	**3x3**	**3x3**	**9x9**
avg. area difference (km²)	86.38	74,71	111,17	94,12
avg. Hausdorff diff. (km)	8,90	9.03	8.93	9.02
avg. cent. point diff. (km)	0,32	0,26	0,26	0,25
avg. perimeter diff. (km)	125,9	113,6	116,1	119,3

Table 1: Evaluation of the similarity of the integrated data with the reference data by evaluating different parameters.

A reason for these not satisfying results are that the used quality measures are on the one hand somehow correlated and on the other hand the consideration of a single measure is not sufficient to describe the quality, because one of the measures may indicate high similarity, but another measure may indicate low similarity at the same time. What we need is an integrated quality measure that considers different geometric aspects at the same time. However, an integration is not trivial, because the measures cannot just simply be added. A solution could be a statistical approach like it is described in Walter and Fritsch (1999). In this approach, measures from the information theory are used to integrate different similarity measures for the matching of spatial data.

When we look at the input data, it can be seen that some of the collected polygons have very low quality (see Figure 7). This has a negative effect to the quality of the integrated data. If we can detect those polygons automatically, we can remove them before the integration and we can expect that the quality of the integration will increase. In a test, we evaluated how the similarity measures of the multiple representations are statistically distributed (see Walter and Sörgel, 2018).

Figure 8 shows the frequency distribution of the area difference of a polygon that was collected 56 times by different students compared with a reference polygon. The distribution shows an approximated normal distribution and the maximum frequency is near that value where the similarity difference between the collected objects and a reference object is minimal (difference = 0). If this is the case, we can identify automatically the polygons that are very dissimilar and remove them before the integration. We found similar results also for the other similarity measures.

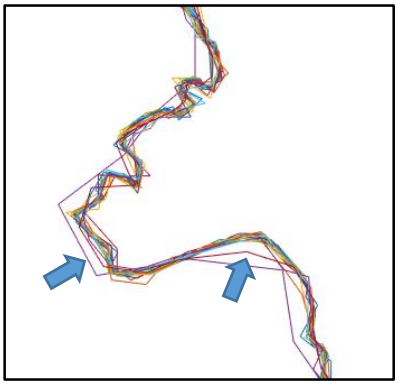


Fig 7 Polygons collected with low quality

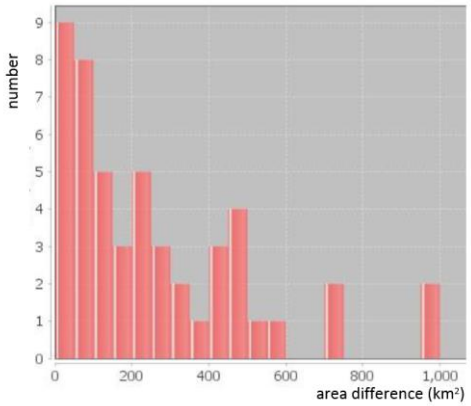


Fig 8 Frequency distribution of the area difference between multiple collected polygons and a reference polygon

6. DISCUSSION

In this paper, we have discussed a raster-based approach for the integration of multiple collected vector data. One of our ideas was that we could utilize the Wisdom of the Crowd (If many individuals measure the same object, the average geometry should be near the real geometry). This idea has proved only partially. The integrated geometry is better than many of the individual representations but small details can disappear since there is inherent smoothing effect because of the multiple representations.

However, these are first results and there is room for improvements. For example, it is easy to identify those representations, which are very dissimilar to the other representations. Typically, these representations are inaccurate because most of the representation are indeed near the real geometry. If we remove the dissimilar representations before the integration, we can expect better integration results because we eliminated the “outliers”.

Another point is that we can easily identify areas, where the collection of the objects was difficult because the object boarders were difficult to detect in the image. These are the areas were the multiple representation have very different geometries. Figure 9 shows the two isles Rügen and Usedum which are in the northern part of Germany. Some of the students collected the corresponding polygon including the isles and some not because they were not sure if they belong to the state Mecklenburg-Vorpommern or not. An integration in that area

does not make sense because the integrated geometry will go through the isles. However, situation like this can be identified automatically because of the strongly different representations. This enables also to estimate the quality of the integrated polygons by evaluating how different the input polygons are.

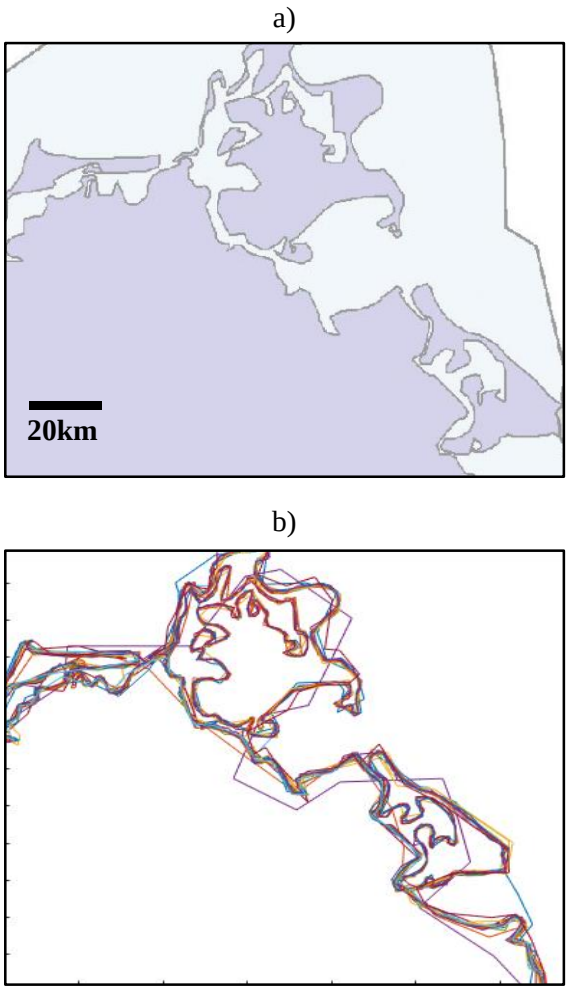


Fig 9 The isles at the boarder of Mecklenburg-Vorpommern were collected by some of the students and by some not: (a) data source; (b) multiple data collection

What we still need is a method to evaluate better the quality of the results. The literature describes many similarity measures to compare spatial data but the consideration of single measures is not sufficient to describe the quality. What we need is an integrated measure. This will be part of our future research.

One positive aspect of our approach is that the computing in the raster-area can be done extremely fast which means that we can easily process also input data that consists of much more different representations as in our test.

REFERENCES

Barron C., Neis P. and Zipf A. 2013. A Comprehensive Framework for Intrinsic OpenStreetMap Quality Analysis. *Transactions in GIS*, 18(6), 877-895.

Bernstein M.S, Little G., Miller R.C,, Hartmann B., Ackerman M.S., Karger D.R. and Panovich, K., 2010. Soylent: a word processor with a crowd inside. In: Proceedings of the 23nd annual ACM symposium on user interface software and technology, 313-322.

Douglas, D.H. and Peucker, T.K., 1973. Algorithms for the reduction of the number of points required to represent a line or its caricature. *Canadian Cartographer* 10(2), 112–122.

Goodchild, M.F., 2007. Citizens as Sensors: The World of Volunteered Geography. *GeoJournal* 69, 211-221.

Lynch, M.P. and Saalfeld, A.J., 1985. Conflation: automated map compilation - a video game approach. Proceedings Auto-Carto 7, 343-352.

Shrier D., Adjodah D., Wu W. and Pentland A., 2016. Prediction markets. Technical report, Massachusetts Institute of Technology.

Sultan, J., Ben-Haim, G., Haunert, J.H. and Dalyot, S., 2017. Extracting spatial patterns in bicycle routes from crowdsourced data. *Transactions in GIS*, 21(6), 1321–1340.

Volz S. and Walter V., 2006. Linking different geospatial databases by explicit relations. *Geo-information Science Journa*l 6(1): 41–49.

Walter V. and Fritsch D., 1999. Matching Spatial Data Sets: a Statistical Approach. *International Journal for Geographical Information Science* 13(5), 445-473.

Walter, V. and Sörgel, U. 2018. Implementation, Results and Problems of paid Crowd-based geospatial Data Collection. *Journal of Photogrammetry, Remote Sensing and Geoinformation Science*, in Review.