Mining Multitemporal in-situ Heterogeneous Monitoring Information for the Assurance of Recorded Land Cover Changes

Kevin Alonso, Daniela Espinoza-Molina, *Member, IEEE*, and Mihai Datcu, *Fellow Member, IEEE*
German Aerospace Center (DLR), Remote Sensing Technology Institute (IMF), Wessling, 82234 Germany

***Abstract*—We present a data mining methodology to filter and validate land cover change detections obtained from multitemporal in-situ surveys. As in-situ data we use the measurements from the European Land Use and Coverage Area frame Survey (LUCAS), which provides images with standardized metadata about land cover and land use within the whole territory of the European Union. Multitemporal LUCAS surveys present an anomaly in the amount of land cover changes that disagree with the estimated by experts. Therefore, our methodology analyses the available data in order to explain the existing irregularities in them. The initial step of our methodology is based on database query refinements. The data mining methodology continues with an image analysis process. This analysis calculates similarity measures of the multitemporal images which are used to identify the potential misclassifications. The final step involves a Geographic Information System (GIS) based on web technologies. By defining different color codes assigned by the similarity measures, the system represents the examined points on a digital Earth globe. There, a user can easily discriminate potentially misclassified points for subsequent detailed analysis or corrections. The final output of the methodology shows remarkable results for detecting misclassified land cover changes.**

***Index Terms*—Big Data, Data Integration, Data Mining, GIS, In-Situ Data, LUCAS, Multitemporal Change Detection.**

I. Introduction

For the last decades, Earth Observation (EO) data have evolved in quality, quantity and heterogeneity. Due to the increasing heterogeneity in the data there are several research works focusing on data fusion and integration. Implementations integrating different EO data sources have been presented for security and hazard decision makers like GEODec [1], or the system introduced in [2] which aims to support earthquake research and disaster response. We can find projects like EOLib [3] or TELEIOS [4] where EO image metadata and linked data are used as query parameters in order to improve EO image retrieval results. Data fusion from third party sources has also been used in [5], where besides the information extracted from EO image analysis data, another layer extracted from OpenStreetMaps was used in the learning stage of the retrieval system. Regarding the use of in-situ data, the work in [6] integrates distributed in-situ data with very high resolution optical EO images providing: geospatial data queries, on demand image processing, and fast map visualizations to support collaborative and more efficient emergency response. Finally, in [7] a web based GIS for visual land cover analytics was presented. The system architecture integrated in-situ data from the European Land Use and Coverage Area frame Survey (LUCAS) [8] with different EO products, e.g., Synthetic Aperture Radar (SAR) images for German satellite TerraSAR-X. This system is a tool that supports EO analysts and expert users by means of information integration or through analytical processes with the final goal of promoting the use and improving the understanding of the EO data.

Every three years since 2006 a survey campaign has been carried out to monitor the state and change dynamics in land use and cover in the European Union (EU) called the LUCAS survey. The survey comprises ground observations that can be divided into three types: 1) micro data of the land cover, land use and environmental parameters associated to the single surveyed points; 2) in-situ photos of each point and landscape photos in the four cardinal directions; and 3) statistical tables with aggregated results by land cover/use at geographical level. In the LUCAS 2009 survey 234,561 points were visited in-situ by 500 field surveyors on 23 countries, defining 77 different land cover classes. The LUCAS 2012 survey includes 270,389 points visited in-situ by 594 field surveyors on 27 countries, defining 83 different land cover classes. In 2015, between March and October,the LUCAS 2015 survey took place. Surveyors from 28 Member States visited a total of 273,401 points.

The amount of information collected until now reaches tens of Terabytes. This volume of information is already big enough to make it impossible for the data to be supervised at small scale. In consequence, the task of collecting and supervising the data relies on the big group of field surveyors.

Analyses of the acquired multitemporal data have shown a very high variability in the land covers, exceeding the expectations of the experts. Using this peculiarity as motivation we started a deeper analysis of the LUCAS surveys aiming to identify the real land cover changes from the potential inaccuracies introduced during the recording or annotation of the data. Table I shows two LUCAS survey points. While the first point shows a clear example of land cover change, the other one shows a point with a

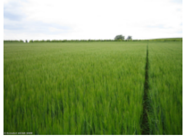
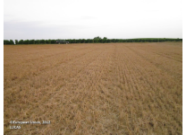
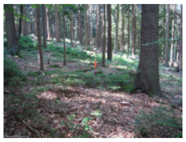
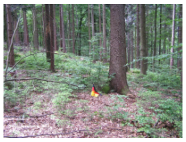
	2009	2012
Real Land Cover Change		
Uncertain Land Cover Change		

TABLE I

Multitemporal LUCAS survey points. The first point shows a clear example of land cover change. In contrast, the second point presents a non visible land cover change. In the LUCAS survey records both points contain land cover changes. In this way, the second point could be considered as an example of the existing irregularities.

non-visible land cover change. In reality, both points are marked as land cover changes. In our understanding, the latter point is an example of the aforementioned anomalies, where the recorded land cover change is uncertain.

With this study we aim to provide tools for the quality assurance of the existing and future LUCAS surveys. Furthermore, we aim to improve the impact and integration of the in-situ observations in EO applications. Consequently, this paper presents a data mining methodology to filter and validate land cover change detections obtained from multitemporal LUCAS in-situ surveys.

The rest of the paper is structured as follows: Section II presents the architecture of the system. Section III explains the data mining methodology which is composed of three steps: Section III-A describes the data refinement step, Section III-B the image analysis and similarity computation step, and Section III-C presents the final step which comprises data visualization and filtering processes. After the methodology presentation, each step is evaluated independently in Section IV, Section V, and Section VI, respectively. The evaluation continues in Section VII extending it to the whole methodology. Final conclusions are presented in Section VIII.

II. Data Mining System Architecture

The base of the presented mining methodology is the system for heterogeneous geospatial data analytics first presented in [7]. The system, as shown in Fig. 1, follows a server-client philosophy. The server side is composed of four different layers: (1) the raw data layer, (2) data ingestion layer, (3) database management system, and (4) user oriented web functionality layer. The raw data layer contains the original information sources that are analyzed and processed in the data ingestion layer in order to pass the obtained information to the database management system layer which will store it in a more accessible way, facilitating the querying and visualization operations performed by the user oriented web functionalities layer. In this last layer, the Geo-Information Visualization block performs the communication protocols with third party geographical information providers using a parallel instance of MapServer [9]. Mapserver centralizes all the communication with the third party providers working as a proxy for the main system which remains isolated and avoids cross-domain communications.

The client is constituted of the Graphic User Interface (GUI) layer accessible from any electronic device capable of running a HTML-5 compatible web browser. A detailed description of the architecture and its functionalities can be found in [7].

In this paper, we will focus on the new or updated system modules represented with a darker color in Fig. 1. The Feature Extraction module is newly introduced in the architecture and it is responsible of the in-situ image analysis processes. The geographical database, the Metadata and Statistical Visualization module, along with the Image Analytics module were already part of the system architecture but their functionalities were upgraded considering the requirements of the presented data mining methodology. The PostGIS [10] is a community developed open source spatial database extender which allows the geographical querying of the spatial data and it is based on PostgreSQL technology [11][12]. The Metadata and Statistic Visualization module collects the data from the database to generate the required data visualization. The Image Analytics module manages the interaction of the user with the data and sends the required instructions to Geo-Information and statistical visualization processes in case an update is required. In Section III the specific functionalities of each module are described in detail.

III. Data Mining Methodology

A preliminary study of the survey methodology shows small improvements in the survey protocol with the pass of the years. One remarkable change is the increase number of the surveyed points. Consequently, different amounts of information are available in the database for each surveyed point. Although it is important to be aware of this fact, it does not impact the change detection procedures. A second change, the one that could explain at a certain level the high multitemporal land cover changes, is the update in the land cover hierarchical class structure between the surveys done in 2009 and 2012. The multilevel hierarchy starts with general land cover classes at the lower level and extends to more specific classes with each higher level. The hierarchy changes were limited to the inclusion of third hierarchical level classes inside the first level classes *Woodland* and *Bare Land*. Nonetheless, even this hierarchy structure change solely has produced a non-realistic increase in the detected changes. In order to quantify the non-realistic land cover changes due to the hierarchy modifications and detect other possible misclassification sources we introduce a data mining methodology which comprises three different steps: 1) database query refinement, 2) in-situ image analysis and similarity measure computation, and 3) on map data visualization and filtering.

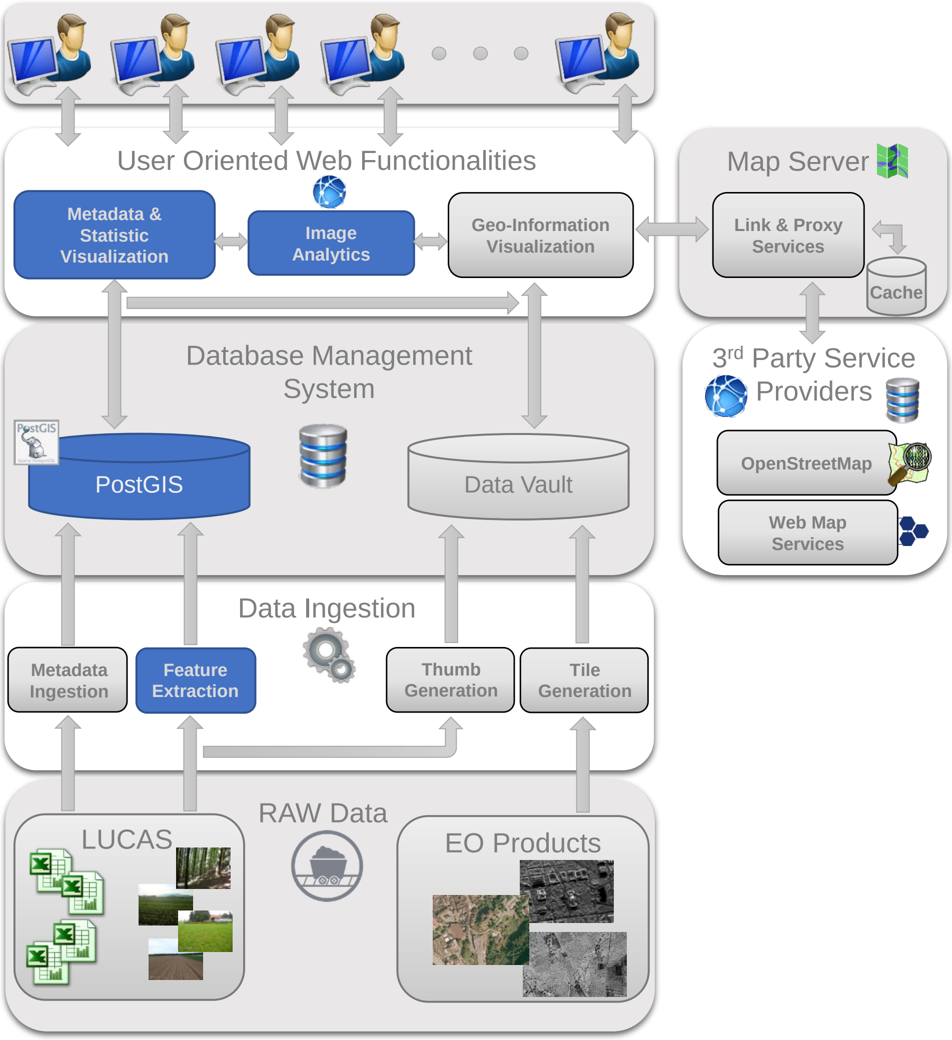


Fig. 1. System architecture. Following a server-client philosophy, the system is designed to rely all the computational complexity over the server making the client side lightweight.

A. Database Query Refinement

The data mining methodology for land cover change detection starts with the mapping of the hierarchical structure changes. The objective of the mapping is to exclude the points that only present land cover changes due to the land cover class hierarchy modification. The changes are shown in Table II. The changes were limited to only two of the first hierarchical level classes *Woodland* and *Bare Land*. Woodland's second level hierarchy members *Coniferous Woodland* and *Mixed Woodland* were extended with a third hierarchical level formed by: *Spruce Dominated*, *Pine Dominated* and *Other*. In 2009 *Bare Land* was just defined as a first hierarchical level, but in 2012 it's definition was extended with a second level hierarchy formed by: *Rocks and Stones*, *Sand*, *liches and Moss* and *Other*.

The mapping of these hierarchy changes is implemented at the Metadata and Statistic Visualization module. There, it is possible to refine the database requests in Structured Query Language (SQL) [13]. By mapping the land cover classes affected by the hierarchy changes. As a result, the queries to the database requesting points containing land cover changes will exclude the points whose land cover change was among the mapped ones, and hence it will avoid the introduction of false-positives.

B. In-situ Image Analysis and Similarity Measure Computation

After the query refinement, the methodology exploits the available point images. Our integration of the Feature Extraction module in the architecture offers valuable new data obtained from image analyses which extend the LUCAS information. In this way, our PostGIS database is extended to link and store the results from two different image analyses. The first one corresponds to the Bag-of-Words (BoW) [14] generated using a common dictionary of RGB colors. The second analysis extracts at image level the texture information using the Weber's Law Descriptor (WLD)[15].

Being D a given image for analysis, the first image

Common Hierarchy		Hierarchy Update 2012
Woodland	Broadleaved	-
	Coniferous	Spruce Dominated Pine Dominated Other
	Mixed	Spruce Dominated Pine Dominated Other
Bare Land		Rocks and Stones Sand Lichens and Moss Other

TABLE II

Land cover class hierarchy modifications between 2009 and 2012 LUCAS surveys. The Woodland second level hierarchy members *Coniferous Woodland* and *Mixed Woodland* were extended to a third hierarchical level formed by: *Spruce Dominated*, *Pine Dominated* and *Other*. Furthermore, the first hierarchical level *Bare Land* was extended with a second level hierarchy.

processing step results in an image color quantization based on a predefined color map. The assigned color map is created by dividing uniformly the color space in 256 elements. Once the D is quantized the BoW is generated defining $p(\omega_{RGB}|D)$ as the probabilities of the words in a given image. Where ω_{RGB} represents the 256 words in the dictionary and the index RGB the identifier of each word.

The second analysis takes again an image D as input for our WLD algorithm that generates as output a WLD histogram that we will use as a second BoW dictionary, $p(\omega_{WLD}|D)$. In this case the words ω_{WLD} represent the different combinations of excitation levels and orientations taken into account in the WLD algorithm. In our implementation we decided to use 18 different excitation levels and 8 orientations.

The described analyses procedures are done over all the available LUCAS images. The obtained results are stored in a database, linking the source images and the corresponding surveyed point. After the information extraction and analysis phase comes the result ranking process. This process is implemented in the Metadata and Stastic Visualization module following a classical approach used by the image analysis community. We use the extracted feature sets as visual signatures to compute similarities among images. Examples of these uses can be found in [16] and [5]. In this specific scenario we use the stored $p(\omega_i|D)$ to compute similarity distances and generate a ranking of the LUCAS points, where i is the index indicating the length of the used dictionaries. The available dictionaries are the previously described $p(\omega_{RGB}|D)$, $p(\omega_{WLD}|D)$ and the joined dictionary obtained from the concatenation of these two $p(\omega_{RGB}|D) \cup p(\omega_{WLD}|D)$.

For the calculation of the similarities of the images from the multitemporal survey we decide to use Kullback-Leibler distance

$$d = \sum_i p(\omega_{i_{2009}}|D) \cdot p(\omega_{i_{2012}}|D) \cdot ln\left(p(\omega_{i_{2009}}|D)\right) \qquad (1)$$

where $p(\omega_{i_{2009}}|D)$ is the probability vector of a given image from 2009 survey and $p(\omega_{i_{2012}}|D)$ it's equivalent in the 2012 survey.

As mentioned in the introduction, each LUCAS point is composed of five different images. The main image shows the exact GPS (p)oint that has been surveyed and the other four photos cover the surroundings of the location by showing the different cardinal directions: (N)orth, (E)ast, (S)outh, and (W)est. Here, we define each LUCAS point as a five element vector P,

$$P = [d_p, d_N, d_E, d_S, d_W] \qquad (2)$$

where the distances d_* are computed according to (1) for each pair of multitemporal images available in the original LUCAS point.

For the sake of posterior visualization simplicity we aim for an unique similarity value for each LUCAS point. Thus, we calculate a similarity value as a weighted mean of the elements in P. To that end, we define a weight vector

$$W = [w_p, w_N, w_E, w_S, w_W] \qquad (3)$$

where the different weights w_* are assigned as follows: $w_p = 0.3$ given that it always contains the analyzed land cover; and the remaining weights are set to 0.175 in contemplation of possible changes of the surroundings which also affect the analyzed land cover. The similarity can then be computed as

$$Similarity = \frac{\sum_{n=1}^{5} w_n \cdot d_n}{\sum_{n=1}^{5} w_n}. \qquad (4)$$

Finally, by means of the calculated point similarity values it is trivial to generate a ranking, listing the points in order of image similarity between their images among surveys.

C. On Map Data Visualization and Filtering

The implementation of the on map visualization and filtering procedures requires modifications in the user oriented web functionalities layer. The Metadata and Statics Visualization module capabilities are extended in pursuance of a better and faster visual discrimination of the data differences. The approach followed includes a redefinition of the markers used to represent the survey points. The markers implement a color codification which allows us to represent the results obtained from the similarity rankings which we use as confidence value of the annotation. Additionally, we are able to include another color indicators to visually represent the time span and the distance between the survey acquisitions. The time of the year when the information of the point was collected in each survey can be meaningful to explain some of the dissimilarities in the images.

Regarding the Image Analytics module, different UI elements have been introduced in order to help the data filtering process and the annotation modification. Furthermore, new server-client and inter-module communication processes have been implemented to support the new functionalities required. Fig. 2 shows part of the user interface during the visualization and filtering step. The points shown are the result of the previous two steps. In this specific case the color intensity from brighter to darker indicates the similarity between the images of the multitemporal survey points. The black color indicates high similarity while the brighter tonalities represent lower similarity. The slider-bar over the map can be used to filter the points drawn over the map. It can set different distance thresholds which are used in the querying process as condition that the points must fulfill in order to be retrieved. The slider-bar has three different operation modes. It can retrieve the points: 1) under the threshold, 2) over the threshold, and 3) in the range between to thresholds. The specific point information is visualized individually by showing a table summarizing the most relevant data (under the map frame), and the acquired multitemporal images (right). The multitemporal images are group by survey year. The top image corresponds to the top the exact GPS point surveyed and it is followed by the images pointing to the different cardinal directions, starting with the north and continuing clock-wise. The selection of a specific point can be done in three different ways. First, by selecting the markers over the map. Second, by using the buttons in the last column of the information table. And third, by using the pagination roulette at the bottom of the UI. This roulette indexes all the points represented on the map.

IV. Query Refinement Evaluation

For the evaluation of the query refinement processes we use the LUCAS data of Germany and Spain from the surveys of 2009 and 2012. A summary of the evaluation is presented in Table III.

A. Case Study Germany

Germany's LUCAS data sum a total of 46084 surveyed points. The initial points are reduced to the points that share the same geolocation, 37504. The difference in points, as explained in Section III, is due to the increase of the surveyed points in the 2012 survey. Hence, we have a total of 18752 pairs of points in which we can perform the temporal land cover change detection. Querying only by the change on land cover will return a total of 9240 geographical points with land cover changes, the 49.35% of the total. The query refinement, described in III-A, implements the modifications of the land cover hierarchy and detects 2496 points annotated as land cover changes, the 13.31%, which should not be marked as land cover change. These points can be differentiated by the hierarchy change type. Thus, an 8.67% correspond to *Coniferous Woodland* hierarchy change, a 4.43% to the *Mixed Woodland* change and a 0.21% to the *Bare Land* change. Examples of the points that are discarded are shown in Table IV. At this stage, the number of points with possible land cover changes has been reduced to 6744, a 35.96% of the total. Comparing the number of points discarded with the originally annotated as change, we can say that in Germany the 27.01% of the detected land cover changes were related to the modification of the class hierarchy and not real land cover changes.

B. Case Study Spain

The LUCAS data for Spain contains a total of 65290 surveyed points. The number of multitemporal points available is 25016. The initial query for the retrieval of the land cover changes returns 12172 points, the 48.66% of the total. After the query refinement procedures we can discard 2020 points, the 8.07%. Looking at the hierarchy change type, the 5.43% corresponds to *Coniferous Woodland* hierarchy change, a 1.60% to the *Mixed Woodland* change and a 1.04% to the *Bare Land* change. Therefore, the number of points with possible land cover changes can be reduced to 10152, a 40.59% of the total. In this case, the number of points discarded versus the initially annotated with land cover change is the 16.60%.

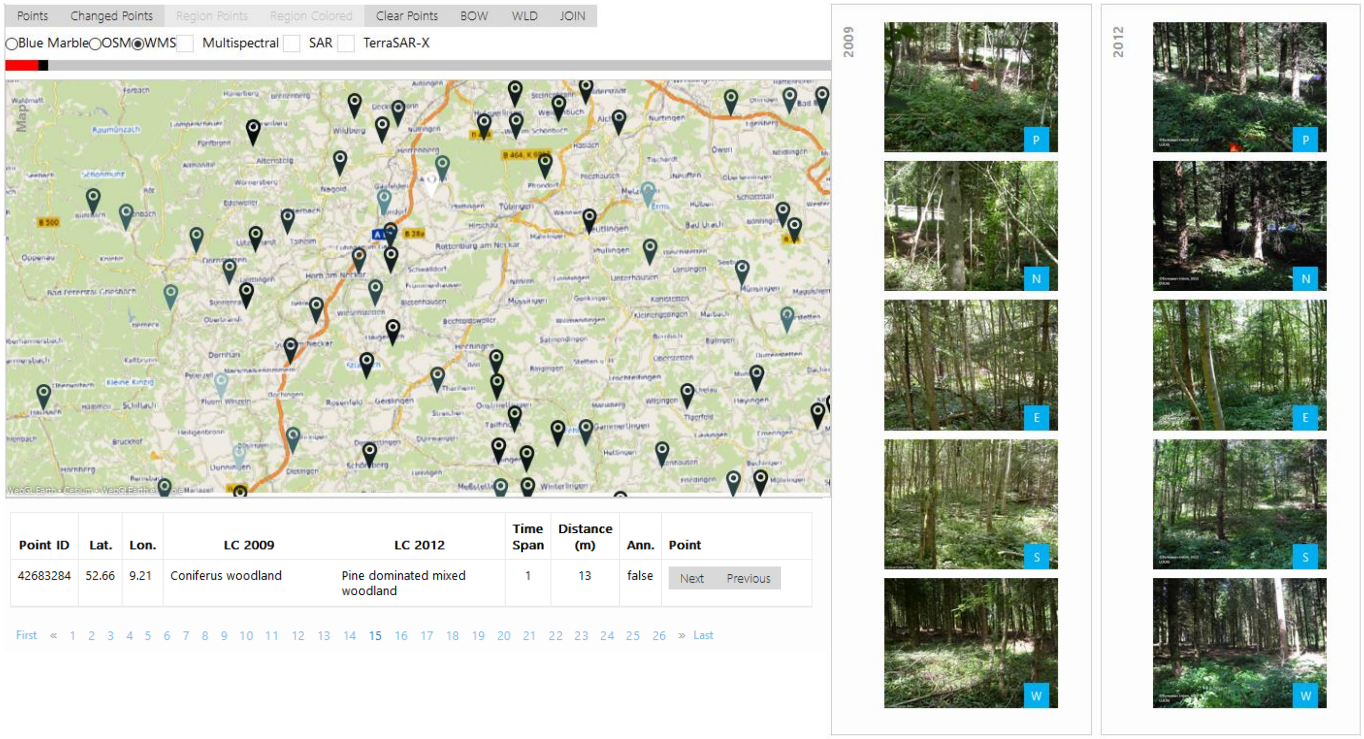
V. Image Analysis and Similarity Measure Evaluation

The methodology's second step generates the similarity measures used for ranking and color coding the multitemporal points. The system implements the possibility to generate three different rankings based on the analysis described in Section III-B. The rankings obtained by using the RGB dictionary and the join dictionary present a better results comparing with the WLD dictionary. The ranking of the former two shows a higher visual coherence clearly ranking the most similar and dissimilar point at the extremes of the ranking. The ranking obtained with the WLD dictionary presents less consistent results, interleaving high similarity points with low similarity ones.

Table V presents the information of different points with high and low similarities using the join dictionary. The first four examples correspond to the high similarity values. High similarity points are marked as low certainty of containing land cover changes. While most of the high similarity points annotated with multitemporal change do not seem to have any land cover change, there are some of the points, e.g. middle-left point in Table V, that even showing a high similarity also contain a land cover change. The last row shows points with low image similarity. At this side of the ranking, the majority of the points correspond to agricultural lands where the change in the crop type is clearly visible. These points with low image similarity are the ones that should be marked with the higher certainty of land cover change.

VI. Data Visualization and Filtering Evaluation

At this point of the methodology we will focus our evaluation on the usability of the developed tools for the data visualization, filtering and correction. To evaluate the performance of the complete methodology and the usability of the developed tools we focused our analysis



Point ID	Lat.	Lon.	LC 2009	LC 2012	Time Span	Distance (m)	Ann.	Point
42683284	52.66	9.21	Coniferus woodland	Pine dominated mixed woodland	1	13	false	Next Previous

Fig. 2. User interface used during the step three of the data mining methodology. The tool allows the visualization of the degree of reliability of the detected land cover changes. It also implements capabilities of filtering, selection and correction of annotations.

	Germany		Spain	
	Number of Points	Percentage	Number of Points	Percentage
Multitemporal Points	18752	100	25016	100
Annotated Changes	9240	49.27	12172	48.66
Detected Hierarchy Changes	2496	13.31	2020	8.07
Coniferous Woodland	1625	8.67	1360	5.43
Mixed Woodland	832	4.43	400	1.60
Bare Land	39	0.21	260	1.04
Changes After Refinement	6744	35.96	10152	40.59
Discarded vs. Annotated	-	27.01	-	16.60

TABLE III

Evaluation of the query refinement step using LUCAS data of Germany and Spain from 2009 and 2012. The annotated land cover changes in both cases are around 50%. Applying the query refinement procedures to filter the class hierarchy modifications between 2009 and 2012 LUCAS surveys the points discarded as land cover change are 13% for Germany and 8% for Spain. The ratio of the discarded land cover changes relative to the annotated changes is of a 27% in Germany and of a 16% in Spain.

in the LUCAS data of Germany. At his point the user will exploit the outputs of the previous methodology steps. The initial step provided the query refinement, as described in Section IV-A, where the number of points containing land cover changes was reduced a 27.01%, from the initial 9240 to 6744. The second step generated the similarity ranking of the points that are used here to represent the confidence level of the land cover change annotation.

The developed tools allowed a user to review effectively the remaining points in around one working day. At the end of the data visualization and filtering process all the points were reviewed. The land cover changes of the 72.2% of the points were validated. The multitemporal changes of the other 27.8% were discarded.

VII. Evaluation of the Data Mining Methodology

After evaluating independently each of the methodology's steps we proceed to evaluate the performance of the entire methodology. It is clear to us the query refinement step is a valuable process in order to initially narrow down the number of points to be analyzed. The image analysis

Year	Land Cover	Point Images				
		Point	North	East	South	West
2009	Coniferus Woodland					
2012	Other Coniferous Woodland					
2009	Mixed Woodland					
2012	Spruce Dominated Mixed Woodland					
2009	Bare Land					
2012	Other Bare Soil					

TABLE IV

Examples of the points affected by the land cover hierarchy modifications: *Coniferous Woodland*, *Mixed Woodland*, and *Bare Land*. Similar points are discarded during the query refinement step of the methodology because they do not contain a real land cover change.

and similarity computation step can be re-evaluated by using the results obtained in the evaluation of the third step in Section VI. Thus, using the corrected annotation of the land cover changes as a ground truth, a quantitative analysis of the quality of the similarity ranking for land cover change detection can be computed. In other words, we use the validated results of the land cover changes, obtained from the data mining methodology, to measure the performance of similarity rankings for detecting real/non-real land cover changes. The analysis is performed using the ranking generated by the RGB dictionary. For the case of retrieving the points with a real land cover change, the obtained results are presented in Fig. 3a. Here, there are represented the precision, recall, accuracy and F_1 measures, i.e, the equally weighted harmonic mean of precision and recall. A detailed explanation of the used measurements can be found in [17]. The results show very high precision values, over 90%, when limiting the ranking up to 2000 points. The accuracy and $F1$ parameters start at lower values but they values increase with the amount of retrieved images and the improvement of the recall. While retrieving around 5000 points the maximum performance is offered obtaining a precision of 83.67%, a recall of 81%, an accuracy rate of 74.8% and a $F1$ value of 82.3%.

On the contrary, if we inverse the approach and analyze the performance of the ranking for retrieving the points with a non-real land cover change, the obtained results are not so promising. The Fig. 3b shows how the precision value rapidly decays to 75% when limiting the retrieved points to 500. When trying to retrieve the same amount of points annotated as non-real land cover change, 1875, the precision value is just a 55.25% with a recall of 55.9%, an accuracy of 75.27% and a $F1$ value of 55.57%. These poor results are explained by the points similar to the one shown in Table V, where even having high visual similarity, the land cover changes exist.

In our opinion the results obtained with the similarity ranking for the case of low similarity points could offer good enough results for some kind of automatization. On the other hand, the results with the high similarity points are not good enough. Hence, we think the data visualization and filtering tools play an important role in the proposed data mining methodology. This last step uses the generated ranking result in order to facilitate the user

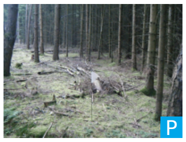
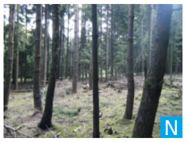
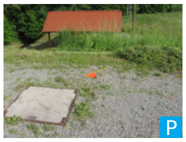
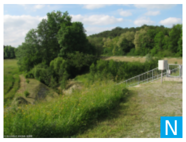
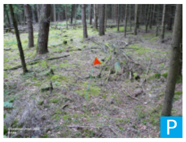
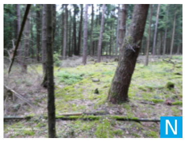
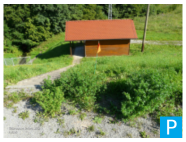
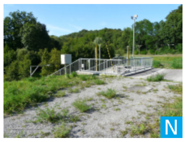
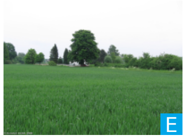
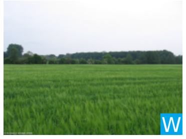
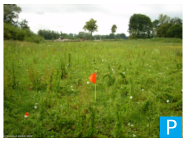
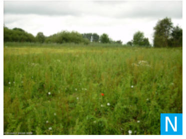
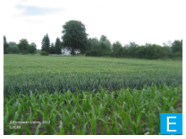
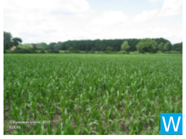
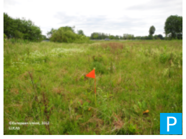
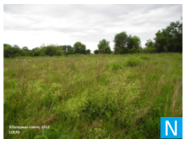
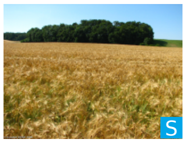
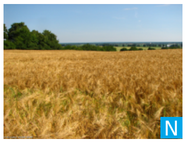
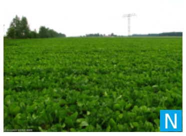
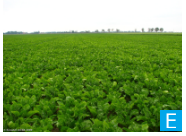
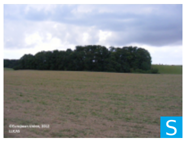
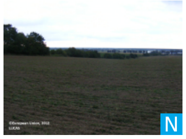
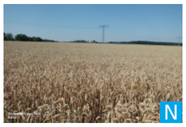
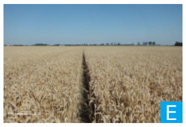
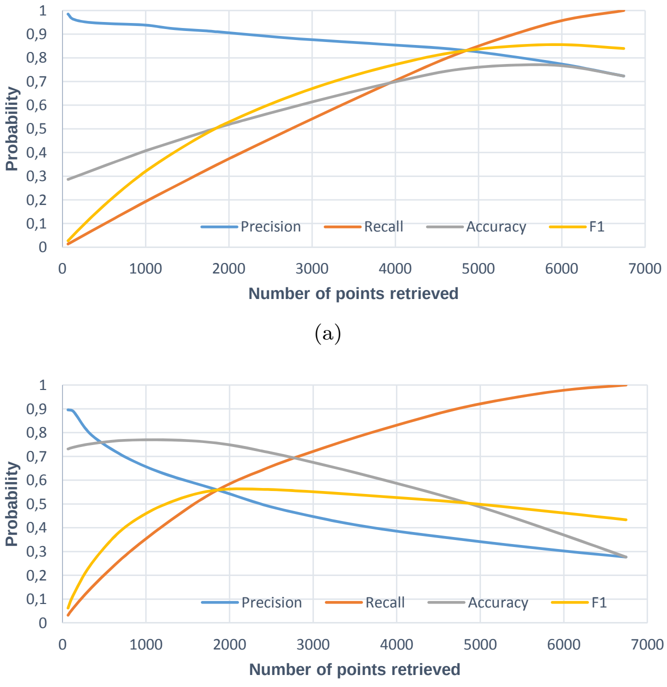
Similarity	Year	Land Cover	Point Images		Land Cover	Point Images	
High	2009	Coniferous Woodland			Non built-up Linear Features		
	2012	Spruce Dominated Mixed Woodland			Grassland without Tree Cover Woodland		
		Distance	0.13	0.16	**Distance**	0.43	0.64
	2009	Barley			Grassland without Tree/Shrub Cover		
	2012	Maize			Spontaneous Re-vegetated Surfaces		
		Distance	0.18	0.20	**Distance**	0.25	0.26
Low	2009	Barley			Sugar Beet		
	2012	Spontaneous Re-vegetated Surfaces			Common Wheat		
		Distance	11.04	11.37	**Distance**	10.38	10.00

TABLE V

Image analysis and similarity measure evaluation. The points shown correspond to the opposite extremes of the ranking using the join dictionary described in Section III-B. High similarity points are marked as low confidence annotation points to contain a land cover change. We can appreciate that most of the high similarity points don't really contain a real land cover change. On the other hand, the annotations of low similarity points are marked as highly reliable considering most of them contain land cover changes.

task of reviewing, filtering and correcting the land cover annotations. The entire process can be performed with an affordable time investment, offering a better final results that any possible automatization. Thus, at the end of the three steps of the data mining methodology we reduce the original 49.27% of the points annotated as land cover changes, to only the 25.97%.

Additionally, the data visualization tools have helped to identify and understand the most common land cover misclassifications. Some of the examples of the errors are listed in Table VI. One of the most common misclassification mistakes are related to the linear features, i.e., roads. The first case shows a point of a road inside a forest. The landscape did not change but in 2009 the surveyors decided to classify it as *Non Built-up Linear Feature* while in 2012 they preferred to focus more on the surroundings assigning the point the *Broadleaved Woodland* class. Another common mistake is the one shown in the second case where the criteria for the definition of a *Non Built-up Area Feature* or *Non Built-up Linear Feature* is not totally clear. Case 3 and Case 4 show errors in the classification due to small distance differences in the surveyed points. We have also noticed different criteria when classifying grass fields. Specifically troublesome appear to be the land covers *Grassland without Tree/Shrub Cover*, *Grassland with Sparse Tree/Shrub Cover*, *Temporary Grassland*, and *Spontaneous Re-vegetated Surfaces*. Examples of these class misclassifications are shown in the Cases 5 to 7. *Inland Running Water* class also appears to have classification problems, see Case 8. Here, as in previous cases, the error is due to small distance differences between the points. Some less common classification errors include *Apple Tree* and *Cherry fruit* classes. Finally, there are two common misclassifications that appear to happen inside residential areas. In the first one, the surveyors usually



(a)

(b)

Fig. 3. Quantitative results for the retrieval of points using the similarity metric for retrieving: (a) real land cover changes and (b) non-real land cover changes.

change the classification criteria. In the initial survey they decided to annotate a building while in the second survey they decided to annotate the garden of the building. The second classification error in residential areas involves the classes *Buildings with 1 to 3 Floors* and *Buildings with more than 3 Floors*, example of this error is shown in Case 10.

VIII. Conclusions

We have presented a data mining methodology that is able to successfully filter false land cover changes from the real land cover change detections in multitemporal LUCAS in-situ surveys. We shortly introduced the heterogeneous geospatial data analytics system [7], which is the base platform of the presented data mining methodology. We have described the three methodology steps and evaluated them independently. The database query refinement step maps the changes in the class hierarchy in order to exclude the points that only present land cover changes due to the hierarchy's modification. The evaluation of this step has shown relevant land cover change filtering capabilities. The query refinement was able to discard the 27.01% of the data annotated as land cover changes in Germany and a 16.6% in Spain. The second step, the image analysis and similarity computation step, showed big capabilities generating similarity rankings with the point images. In the third step, the visual evaluation of the ranking is very good. It clearly positions at the extremes the most similar and dissimilar points. Unfortunately, a correct ranking based on similarity does not ensure a good discrimination of the land cover changes. This fact can be seen in the quantitative analysis performed in Section V where the dissimilar images offer a very good land cover change detection but failed to detect the changes in more similar images. The data visualization step takes advantage of the previous results in order to offer simplicity and efficiency to the users in their data reviewing tasks. The developed tools allow the reviewing task with a small investment in manpower and time. The final data mining results show a clear reduction in the total number of land cover changes which go from the initial 49.27% to only the 25.97%. Additionally, the data mining methodology has improved our knowledge of the data and has helped us to identify common mistakes done during the surveying campaigns. In this sense, future developments will exploit the methodology results via interactive graphical visualizations which would allow a better comprehension of the mistakes done on every specific land cover change. In our understanding the final quality of future surveys could be improved in two different ways. First, the surveyor training could be improved by presenting the detected common mistakes during the training sessions. Second, the developed system can easily be accessible to the surveyors on the field which will provide fast information of the previous surveys reducing the uncertainty and the subjective criteria in the decision making process.

IX. Acknowledgment

The authors would like to thank the European Commission for providing the metadata and images of the LUCAS surveys.

References

[1] C. Shahabi, F. Banaei-Kashani, A. Khoshgozaran, L. Nocera, and S. Xing, "GeoDec: A Framework to Effectively Visualize and Query Geospatial Data for Decision-Making," *IEEE Multimedia*, vol. 17, no. 3, pp. 14–23, 2010.

[2] J. Wang, M. Pierce, Y. Ma, G. Fox, A. Donnellan, J. Parker, and M. Glasscoe, "Using service-based gis to support earthquake research and disaster response," *Computing in Science Engineering*, vol. 14, no. 5, pp. 21–30, Sept. 2012.

[3] "EOlib. Earth Observation Librarian.," Dec. 2011, [Online: http://deepenandlearn.esa.int/tiki-index.php?page=EOLIB+Project].

[4] D. Espinoza-Molina and M. Datcu, "Earth-Observation Image Retrieval Based on Content, Semantics, and Metadata," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, no. 11, pp. 5145–5159, Nov. 2013.

[5] K. Alonso and M. Datcu, "Accelerated Probabilistic Learning Concept for Mining Heterogeneous Earth Observation Images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 8, no. 7, pp. 3356–3371, 2015.

[6] D. Brunner, Student Member, G. Lemoine, Senior Member, F. Thoorens, and L. Bruzzone, "Distributed Geospatial Data Processing Functionality to Support Collaborative and Rapid Emergency Response," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 2, no. 1, pp. 33–46, 2009.

[7] K. Alonso, D. Espinoza-Molina, and M. Datcu, "Multilayer Architecture for Heterogeneous Geospatial Data Analytics: Querying and Understanding EO Archives," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2016, Accepted for publication under revision.

[8] M. Kotzeva, T. Brandmüller, and Å. Önnerfors, Eds., *Eurostat regional yearbook 2014*, European Commission - Eurostat, 2014.

[9] R. Vatsavai, S. Shekhar, T. E. Burk, and S. Lime, "UMN-MapServer: A High-performance, Interoperable, and Open Source Web Mapping and Geo-spatial Analysis System," in *Proceedings of the 4th International Conference on Geographic Information Science*. 2006, GIScience'06, pp. 400–417, Springer-Verlag.

[10] "PostGIS. Spatial and Geographic Objects for PostgreSQL.," 2016, [Online: http://postgis.net/].

[11] M. Stonebraker and L. A. Rowe, "The Design of POSTGRES," in *Proceedings of the 1986 ACM SIGMOD International Conference on Management of Data*, Washington, D.C., USA, 1986, vol. 15, pp. 340–355, ACM.

[12] M. Stonebraker, L. A. Rowe, and M. Hirohama, "The Implementation of POSTGRES," *IEEE Transactions on Knowledge and Data Engineering*, vol. 2, no. 1, pp. 125–142, 1990.

[13] James Groff and Paul Weinberg, *SQL The Complete Reference, 3rd Edition*, McGraw-Hill, Inc., New York, NY, USA, 3 edition, 2010.

[14] Y. Huang, Z. Wu, L. Wang, and T. Tan, "Feature Coding in Image Classification: A Comprehensive Study," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 3, pp. 493–506, Mar. 2014.

[15] J. Chen, S. Shan, C. He, G. Zhao, M. Pietikainen, X. Chen, and W. Gao, "WLD: A Robust Local Image Descriptor," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1705–1720, 2010.

[16] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Image Retrieval: Ideas, Influences, and Trends of the New Age," *ACM Computing Surveys*, vol. 40, no. 2, pp. 1–60, Apr. 2008.

[17] T. Fawcett, "An Introduction to ROC Analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, June 2006.

Year	Case	Land Cover	Point Images		Case	Land Cover	Point Images	
2009	1	Non Built-up Linear Features			2	Non Built-up Area Features		
2012		Broadleaved Woodland				Non Built-up Linear Features		
2009	3	Maize			4	Non Built-up Linear Features		
2012		Grassland without Tree/Shrub Cover				Vineyards		
2009	5	Grassland without Tree/Shrub Cover			6	Temporary Grassland		
2012		Grassland with Sparse Tree/Shrub Cover				Grassland without Tree/Shrub Cover		
2009	7	Spontaneous Re-vegetated Surfaces			8	Broadleaved Woodland		
2012		Temporary Grassland				Inland Running Water		
2009	9	Apple Tree			10	Buildings with 1 to 3 Floors		
2012		Cherry Fruit				Buildings with more than 3 Floors		

TABLE VI

Common misclassification patterns encountered after concluding the data mining methodology over the German 2009 and 2012 LUCAS surveys. Some of the most common misclassification mistakes are related to the linear features, i.e., roads (Cases 1-2). Other errors are due to small distance differences between the surveyed points (Cases 3-4 and 8). Also, grass fields and different type of fruit trees are difficult to classify (Cases 5-7 and 9). Finally, the residential areas have shown common misclassifications (Case 10).