LandSense

A Citizen Observatory and Innovation Marketplace for Land Use and Land Cover Monitoring

Deliverable 5.6

Quality evaluation of citizen-observed data to the LandSense demonstration cases II

European Commission
Horizon 2020
European Union funding for Research & Innovation

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.689812

Project acronym:	LandSense
Project title:	A Citizen Observatory and Innovation Marketplace for Land Use and Land Cover Monitoring
Project number:	689812
Instrument:	Horizon 2020
Call identifier:	SC5-17-2015
Topic	Demonstrating the concept of citizen observatories
Type of action	Innovation action

Start date of project:	01-09-2016
Duration:	48 months

Deliverable number	**D5.6**
Deliverable title	**Quality evaluation of citizen-observed data to the LandSense demonstration cases II**
Deliverable due date	**29-02-2020**
Lead beneficiary	**UNOTT**
Work package	**WP5**
Deliverable type	**Report**
Submission date:	**21.04.2020**
Revision:	**Version 2.0**

Dissemination Level		
PU	Public	x
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium	
CO	Confidential, only for members of the consortium (including the Commission Services)	

Title:
Quality evaluation of citizen-observed data to the LandSense demonstration cases II
Author(s)/Organisation(s):
Gavin Long (UNOTT), Michael Schultz (UHEI), Ana-Maria Olteanu-Raimond ((IGN)
Contributor(s):
Giles Foody (UNOTT), Inian Moorthy (IIASA)

Short Description:
This deliverable completes the work described in deliverable D5.4 (Quality evaluation of citizen-observed data to the LandSense demonstration cases). That report described the LandSense Quality Assurance (QA) platform and provided some examples of its use with citizen-sourced data. The work described in this deliverable gives an extensive account of the large-scale implementation of all QA services across all three LandSense themes (representing six pilot demonstration cases). Detailed results are provided along with interpretation of their impact. The latter leads onto the next, and final, stage of the work, a good practice guide for QA in citizen observatories.
Keywords:
Quality Assurance, Volunteered Geographic Information, LandSense, WP5, image quality, polygon topology, user privacy, categorical accuracy, citizen science, positional accuracy, contributor agreement

History:

Version	Author(s)	Status	Comment	Date
0.1	A-M Olteanu-Raimond	Draft Initial structure		20.12.2019
0.2	G Long	Initial draft	Results for privacy and photo quality QA tools added	20.01.2020
0.3	G Long, G Foody	Revisions and update	Made agreed revisions to draft version 0.2 and added methodology and discussion sections	28.01.2020
0.4	A-M Olteanu-Raimond	Content additions	Added content on pilots in chapter 2 and results for Toulouse pilot	29.01.2020
0.5	G Long	Revisions and update	Finalised draft version of all photo processing content	10.02.2020
0.6	M. Schultz	Added graphs and tables	Combined all versions	20.02.2020
0.7.-08	M. Schultz, A-M Olteanu-Raimond G Long	Revision and combination	Combined all versions jointly during LandSense meeting	21.02.2020
0.9	M. Schultz, A-M Olteanu-Raimond G Long	Content added	Described tables and graphs	27.02.2020
1.1	A-M Olteanu-Raimond	Content added	Contributor agreement sections added	05.03.2020
1.1	G Long	Content added	Added content on positional accuracy and updated photo privacy and quality checks	07.03.2020
1.2-1.4	M Schultz	Content added	Categorical accuracy sections added.	11.03.2020
1.5-1.7	G Long	Content added Structural changes	Finalised content. Revised introduction, added conclusions updated sections and moved content in light of meeting with G Foody	15.03.2020
1.8	G Long, G. Foody	Near Final version	Integrate edits from review	20.04.2020
2.0	G. Long, G. Foody	Submitted version	Final edits and formatting	20.04.2020

Review:			
Version	Reviewer	Comment	Date
0.2	G Foody	Added comments on initial draft version of report for structure and content	21.01.2020
1.1	G Foody	Review of full draft version	09.03.2020
1.7	M Schultz	Review of final draft version	25.03.2020
1.7	L See	QA of final draft version	08.04.2020
1.8	L See	QA of final draft version and approved for submission	20.04.2020

Executive Summary

Citizen observations have the potential to revolutionise the field of Land Use and Land Cover (LULC) monitoring, greatly increasing reporting capacity and enabling near real-time response to emerging environmental hazards (*e.g.*, deforestation, flooding). However, accuracy and other data quality issues are a key concern when utilising citizen observations. A suite of Quality Assurance (QA) tools suitable for LandSense were identified (D5.1) and implemented in phase I of the project (D5.4).

This deliverable reports on the wider implementation of all QA tools with the data collected in phases I and II of the project and provides detailed results from this work. The use and performance of eight QA tools is discussed across all three LandSense themes (urban landscape dynamics, agricultural land use and forest and habitat monitoring) using heterogeneous datasets from six pilot studies. The QA platform performed as designed and no notable operational errors were encountered. Modifications and additions were made to some of the QA tools in light of the findings of D5.4 and are described here. Quality and privacy checks on photographic data collected performed well (e.g. 90%+ accuracy in detecting privacy features) and correlations between photo quality and feature detection were investigated and described. Image blur was not found to be a significant problem and only detected in specific instances (i.e. low light conditions and images taken from moving vehicles).

QA checks were used to assess hundreds of user observations of LULC features and demonstrated the ability to identify areas of both high and low agreement between multiple contributors. Links between contributor agreement and the type of LULC feature are also described. Building on this work, QA checks on the categorical accuracy of user contributions were performed and found to be very promising. It was found that Volunteered Geographic Information (VGI) was of sufficiently good quality for identifying key types of LULC, such as residential land use change or detecting and identifying the urban fabric. However, some specific LULC features were harder for VGI to identify accurately, e.g., distinguishing between different types of agricultural land use. The results outlined in this deliverable will form the basis for development of the LandSense QA good practice guide (D5.7).

Table of Contents

List of Figures

List of Tables

1 Introduction

The considerable potential of citizen-contributed data sets has been widely recognised in many areas of science. In relation to studies of Land Use and Land Cover (LULC), citizens can be an attractive source of Volunteered Geographic Information (VGI) addressing concerns such as the amount, distribution and timeliness of reference data to support analyses of Earth observation data. However, concerns about the quality of citizen-derived data abound. In Deliverable 5.1 (D5.1), some of the means to explore the quality of citizen-contributed data were discussed. These formed the basis of the quality assurance (QA) component of LandSense. A set of QA tools to meet LandSense objectives has been developed. Initial results, based on LandSense pilot case studies, were reported in D5.4, illustrated further potential and informed recent extensions of the tools. This document extends the work reported in D5.4. It uses mainly new data contributed from the LandSense pilots and seeks to demonstrate the use of all the QA tools identified to achieve the LandSense objectives. The core focus is on illustrating the use of every tool using real data acquired in LandSense but not necessarily for every pilot. In total, eight tools (defined in the first column of Table 1) were identified for use in LandSense. Critically, six of the eight tools identified were explored in a preliminary way in D5.4 but here results for all eight are presented. As this is a rapidly developing subject and as data are still being acquired, it is likely that further developments will be made. The latter may help inform good practices for QA and hence the production of D5.7.

A brief summary of the six pilot studies is given in section 2, followed by a description of the QA methodology adopted. Sections 4 and 5 form the principal element of this work, describing and discussing the results based on performing the QA checks on the full suite of pilot data gathered to date by the LandSense project.

2 Data collection

LandSense deliverables D4.5 (urban landscape dynamics), D4.6 (agricultural land use) and D4.7 (forest and habitat monitoring) provide details and a summary of the data collected through the different LandSense pilots. Readers are urged to review these documents if they require detailed information on the pilots and the data collected (citations are provided in the following sections for each pilot). This section is intended to merely provide a brief summary of the pilots to provide context for the discussion of the QA methods and results in the following sections.

Both point and polygon data are collected through the pilots. For the *in-situ* campaigns, photographs are also collected by users. For most of the pilots, multiple contributions exist (*i.e.*, the same location has been visited by more than one contributor). The volume of each data type collected by the pilot studies and processed by the LandSense QA service can be found in Table 1.

2.1 City of Heidelberg – OSMlanduse validation

- **Goal:** Demonstrate how citizen-observed data and Earth Observation (EO) data can update and improve the quality of LULC authoritative databases.
- Data collection process: Four web mapathons were organised in 2018 and 2019. Contributors were asked to validate LU classifications derived from OpenStreetMap (OSM) and remote sensing data (Sentinel 2 imagery). Multiple contributions per location were collected
- Contributors: University students.
- Tools: OSMlanduse and LACO-Wiki.
- LandSense Deliverable: D4.5 (Stickler, 2020).

2.2 City of Toulouse – Paysages pilot

- **Goal:** Evaluate how citizen-observed data can be integrated into the French LULC authoritative database
- Data collection process: eleven web and *in-situ* mapathons were organised in 2018, 2019 and early 2020. Contributors were asked to validate changes detected by the LandSense Change Detection Service (CDS) or current classifications of LULC data (*e.g.,* residential, commercial, industrial). Multiple contributions per location were collected.
- Contributors: public authorities, engineering students, research expert, citizens.
- Tools: Paysages web, Paysages mobile app, and LACO-Wiki.
- LandSense Deliverable: D4.5 (Stickler, 2020).
- Paper in Remote Sensing by Olteanu-Raimond et al. (2020) https://doi.org/10.3390/rs12071186

2.3 City of Amsterdam – MijnPark.NL pilot

- **Goal:** Test new methods for collecting subjective perception and preferences for urban green space and evaluate how this citizen-observed data can be used to influence sustainable urban development.

- Data collection process: one campaign was conducted in 2019. Rembrandt park users were asked to visit highlighted points in the park and give feedback about their feelings about those locations (e.g., noisy, safe, relaxing, etc.). Multiple contributions per location were collected.
- Contributors: citizens.
- Tools: MijnPark.NLmobile application.
- LandSense Deliverable: D4.5 (Stickler, 2020).

2.4 City of Vienna - City.Oases pilot

- **Goal:** Evaluate how citizen-observed data can complement local administrative, urban planning processes by giving information about the usage and perceptions of urban green areas and open spaces.
- Data collection process: six campaigns were conducted in 2018 and 2019. Contributors gave feedback about different locations in Vienna with respect to the usage of the location and their feeling about that location (*e.g.,* secure, clean, attractive, facilities). Multiple contributions per location were collected.
- Contributors: students, citizens, pupils, and school classes.
- Tools: City Oases mobile app.
- LandSense Deliverable: D4.5 (Stickler, 2020).

2.5 Serbia – CropSupport pilot

- **Goal:** Monitor the state of Important Bird and Biodiversity Areas (IBAs) and Natura 2000 sites by focusing on the threats that affect various habitats.
- Data collection process: four campaigns have been conducted during 2018 and 2019. The campaigns consist in validating threats to biodiversity triggered by IBA caretakers and changes detected by the CDS.
- Contributors: IBA caretakers and citizens (CEO volunteers).
- Tools: Natura alert web application and Natura alert mobile application.
- LandSense Deliverable: D4.6 (Mrkajić, 2020).

2.6 Spain and Indonesia - Natura Alert pilot

- **Goal:** Monitor the state of Important Bird and Biodiversity Areas (IBAs) and Natura 2000 sites by focusing on the threats that affect various habitats.
- Data collection process: four campaigns have been conducted during 2018 and 2019. The campaigns consist in validating threats to biodiversity triggered by IBA caretakers and changes detected by the CDS.
- Contributors: IBA caretakers and citizens (CEO volunteers).
- Tools: Natura alert web application and Natura alert mobile application.
- LandSense Deliverable: D4.7 (Capellan, 2020).

3 Methodology

Figure 1 outlines the LandSense quality assurance system (LQAS). The Roman letters I-III represent the flow of events for General Data Protection Regulation (GDPR)-relevant content, which is an essential data collection requirement (DCR), D5.2. Stages I and III occur almost simultaneously, skipping stage II. At its core is the representational state transfer (REST) application programming interface (API) currently hosted on geopedia.world by Sinergise, accessible through the QA section of the LandSense Engagement Platform https://landsense.eu. The REST API is used to manage the LandSense QA system and acts as its primary interaction hub. The interaction of the campaign manager in querying QA reports, the main functionalities and the data are linked via the REST API. A regular flow of events as depicted in Figure 1 occurs in three steps:

(I) The data from contributors captured through a pilot are stored on the respective LandSense partner's server database. Currently, City.Oases, MijnPark.NL and Natura.Alert.App data are stored and protected on Sinergise's geopedia.world and IIASA.at servers, the OSMlanduse validation data and Paysages are stored on their respective institution's servers, while CropSupport data are stored at inosens.rs.

(II) If a given LandSense campaign has been completed or there are sufficient data available on the server, a LandSense manager can login to the LEP and request pilot specific QA reports.

(III) Pilot specific data are called, and relevant functionalities are launched to produce the QA report or augment the respective pilot's data with QA attributes. The functionalities are stored and provided by https://github.com/LandSense/, which is an external service. Libraries related to the polygon check are available through java-language scripts; and contributor agreement is available via R-language scripts and java-scripts. Categorical accuracy, positional offset and positional accuracy are calculated using R-scripts. Any photo-related checks are provided through the Python language scripts.

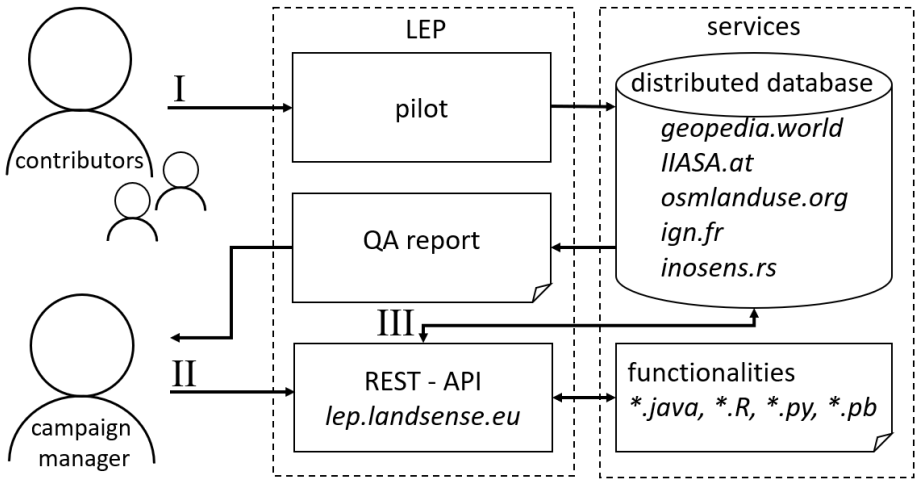

Figure 1: LandSense quality assurance (QA) system schema. The square boxes are the interfaces, the square boxes including dog-ear extension at the bottom right contain modifiable scripts and the dotted boxes refer to concepts. Acronyms and extensions in the figure: LandSense Engagement Platform (LEP), representational state transfer (REST), application programming interface (API),

Java language script file (*.java), R language script file (*.R), Python language script file (*.py), TensorFlow model configuration (*.pb)

The content of a pilot that has been flagged as an essential data collection requirement (eDCR) (D5.2 section 1), such as newly captured imagery, is immediately processed (regardless of a LandSense campaign manager specific query) for GDPR conformity as soon as it is arrives on the server (D5.3 sections and subsections 2.2.1 - 3.2). Relevant artificial intelligence (AI) algorithms provided by external services are controlled by an API (D5.4). Currently, an AI algorithm available through TensorFlow is used for face detection, modified using the Python language. The TensorFlow model configuration contains the learning status and parameters, which can be modified to improve the performance of the imagery-related AI QA services. For license plate detection, a java-based implementation of the OpenALPR (an open source Automatic License Plate Recognition library) was used.

The QA tools, defined in the first column of Table 1, were applied to data acquired from a range of LandSense pilot studies. Table 1 summarises the data used, and a brief discussion of the analyses undertaken for each tool is provided in the sub-sections below. The number in each cell of Table 1 represents the number of contributions processed for that pilot in each QA theme. An "A" denotes the test is applicable but has not been evaluated as part of this deliverable.

Table 1: The set of LandSense data quality analyses applied to different LandSense pilots as adapted from the D5.2 Data Capture Requirements framework.

Theme	**Urban landscape dynamics**				**Agricultural land use**	**Forest and habitat monitoring**
Pilots (institutions)	OSMlanduse validation (UHEI, CHEI)	City.Oases (UBA)	MijnPark.NL (VU)	Paysages (IGN)	CropSupport (INOSENS)	Natura Alert (BLI)
Polygon topology check	-	-		-	202	-
Photo privacy check	-	195	372	260	210	512
Photo illumination check / photo blur check	-	195	372	260	210	512
Positional accuracy	-	878	377	-	207	A
Positional offset	-	443	361	-	211	-
Categorical accuracy	10806	-	-	467	-	571
Contributor agreement	150 sites/points	-	30 sites/points	650 sites/points	-	-

3.1 Polygon topology

Polygon topology verification is only applicable to the CropSupport pilot within the agricultural land use theme. Preliminary topology checks on the initial set of polygon data from this pilot were described in Deliverable D5.4 (Rosser & Schultz, 2019). Readers can also find a complete description of the polygon

topology QA procedure in that report. The same QA check was applied to all polygon data collected since the delivery of D5.4, and the results are provided in section 4.1.

3.2 Photo privacy and quality

QA checks on image data were performed for five of the pilots, representing all three LandSense themes. For the urban landscape dynamics theme, photos collected from the MijnPark.NL, City.Oases and Paysages pilots were processed. Images of agricultural land use from the CropSupport pilot were processed, and 512 images collected by the Natura Alert app were processed as part of the forest and habitat-monitoring theme. In total, 1555 images were processed, with the results of the QA services logged and analysed as outlined in sections 4.2 and 4.3 of this report.

The face detection and blurring tools described in LandSense D5.4 (Rosser & Schultz, 2019) were applied to all image data. Supplementary privacy and quality checks for vehicle license plate detection and photo illumination, as described in LandSense D5.2 (Schultz, 2018), were also implemented and carried out on the image data for all four pilots. The license plate detection method uses the OpenALPR (open source Automatic License Plate Recognition) library[1]. A Java implementation of the library was added to the LandSense QA platform's REST API. For illumination checking, the relative luminance was calculated for each image and used as a proxy for image brightness.

For privacy checks, the primary aim is to minimise omission errors (where the check failed to identify a face or license plate that was visible in the image). Reducing commission errors (where the privacy check falsely identified a feature that is not present in the original image) is the secondary aim of the checks. To this end, the detection thresholds[2] were set to a very low level (0.2 for face detection and 0.1 for license plate detection). These values were determined during the testing phase using pilot data collected in initial data collection phases supplemented by additional photo libraries available. An increased risk of commission error was deemed an acceptable compromise in order to reduce the rate of omission errors.

Thresholds for the photo quality checks (blurring and brightness) were identified using a similar process to those for the privacy checks. Both photo quality checks return an integer value in the zero (Dark/Blurred) to 255 (Bright/Sharp) range. For blurring checks, testing established that any value below 250 on the blurring level showed significant signs of blurring, and the threshold for blurring was, therefore, set to 250. For illumination checks, testing showed that images with brightness levels below 100 were significantly dark and hence, the threshold was set to that level. In order to allow for revisions to quality thresholds as a post process, the values for blurring and brightness levels were also recorded for all images processed.

Results for both photo privacy and quality checks were then manually verified by the QA team. For photo quality checks, this process was relatively straightforward. In the case of privacy checks, where the images clearly contained visible faces/license plates or contained neither feature, the verification was also straightforward. In other instances, features may be present in the image but may not be clearly visible. In

[1] https://github.com/openalpr/openalpr

[2] Detection thresholds are a value between 0 and 1, where lower values lead to higher detection rates but introduce a higher risk of false detections (commission errors).

these instances, the verifier's judgment was used to determine whether and where the feature should have been detected. This issue is further discussed, with examples, in section 4.

3.3 Positional accuracy and offset

Data on the positional accuracy of point-based data from the applicable pilots were collected as outlined in Deliverable D5.2. The positional accuracy is represented by the GPS accuracy of the mobile device used to collect the point data. Values exceeding 20m were excluded from the analysis as outlined in D5.2, Table 6. Analysis of the positional accuracy data for those pilots where GPS accuracy data were provided are given in section 4.4.

Positional offset is defined as the spatial distance between the observer's location and the spatial reference point to which the observation relates. Positional offset was integrated within the pilot's data capture mechanism. For CropSupport, ground observations were valid for association with a related field polygon when they fell within a 25m proximity. For City.Oases, the offset tolerance towards guided points was 30m.

3.4 Contributor agreement

Contributor agreement was assessed on an individual contribution basis as described in Deliverable D5.2, section 3.5. For each contribution, the P_i measure proposed is the computation of Fleiss' kappa coefficient (Fleiss, 1971). Hence, the degree to which the contributors agree on the i^{th} subject, P_i, may be calculated from:

$$P_i = \frac{1}{n(n-1)}\left(\sum_{j=1}^{k} n_{ij}^2 - n_{ij}\right) \qquad [1]$$

where:

- P_i is the contributor agreement for the contribution *i*
- *n* is the number of contributors per contribution, varying from 1 to N
- *k* represents the number of categories
- n_{ij}, represents the number of the contributors assigning case *i* to category *j (*where *j*=1,…, *k*).

Across the pilots, there are two types of contributions: (i) labelling LU or LC classes; and (ii) describing the perception of places and activities in urban areas. Since the latter is subjective, the interpretation of contributor agreement differs from pilot to pilot.

3.5 Categorical accuracy

The categorical accuracy has been assessed when reference data have been available. Usually such data have been captured through validation mapathons. The calculation of categorical accuracy follows consolidated standards as described in D5.2 (section 3.4). In short, classic, key aspects are considered, such as sampling design, response design and analysis (Strahler, 2006). Sampling design and population estimations follows (Foody, 2009), where for sampling size determination, equation [2] is used:

$$n = \frac{z_{\frac{\alpha}{2}}^2 P(1-P)}{h^2} \quad [2]$$

where *n* is the sample size, *h* is the desired confidence interval, P (here, *h* = 0.05) is the respective class proportion of the classification and $z_{\frac{\alpha}{2}}$ is the critical value of the normal distribution for the two-tailed significance level {alpha}. At the 0.05 significance level $z_{\frac{\alpha}{2}}$ equals 1.96.

The difference between the reference data and the map was determined using Overall accuracy (*O*), producer's accuracy (P_j), expressing the underestimation of a class, and user's accuracy (U_i), expressing overestimation of a class, which were calculated and reported. With *q* being the number of classes and p_{ij} the convolution matrix of the classes of the map and the reference data, O is calculated as follows:

$$O = \sum_{j=1}^{q} p_{jj} \quad [3]$$

The class accuracies from the user's and producer's perspective are calculated according to:

$$U_i = p_{ii}/p_{i+} \quad [4]$$

$$P_j = p_{jj}/p_{+j} \quad [5]$$

According to best practices provided by Olofsson (Olofsson et al,2014) accuracy estimates were corrected for the true marginal map proportions (when applicable), substituting p_{ij} by $\hat{p}_{ij}$, correcting for class proportions as matched by the reference data and its samples *n*, where W_i was the estimated proportion for a class *i*:

$$\hat{p}_{ij} = W_i \frac{n_{ij}}{n_{i+}} \quad [6]$$

Accuracy measures were recalculated using equations 2-5 based on the matrix $\hat{p}_{ij}$ corrected for true marginal map proportions. However, the simple accuracy measures of O, P_j and U_i were provided as a baseline, given their intuitive interpretation.

Further details on accuracy assessment and its interpretation, stressing, in particular, the need to consider sampling issues, are given in Olofsson (Olofsson et al, 2014).

4 Results

Results from running the QA tools with the pilot data described in Section 2 are provided here for each theme, as outlined in Table 1. The pilot data processed varies across themes due to the differing QA functionality applicable to the various pilots and the nature of the data collected. The specific nature of the QA processes and the data to which it was applied is given in each theme's subsection.

4.1 Polygon topology check

For the CropSupport pilot, an additional 100 polygons representing fields of crops were collected since the original data collection phase described in deliverable D5.4 (Rosser & Schultz, 2019), resulting in a doubling of the total number of polygons collected. Figure 3 shows the polygons collected from the pilot, with the original polygons shown in yellow and new polygons shown in blue.

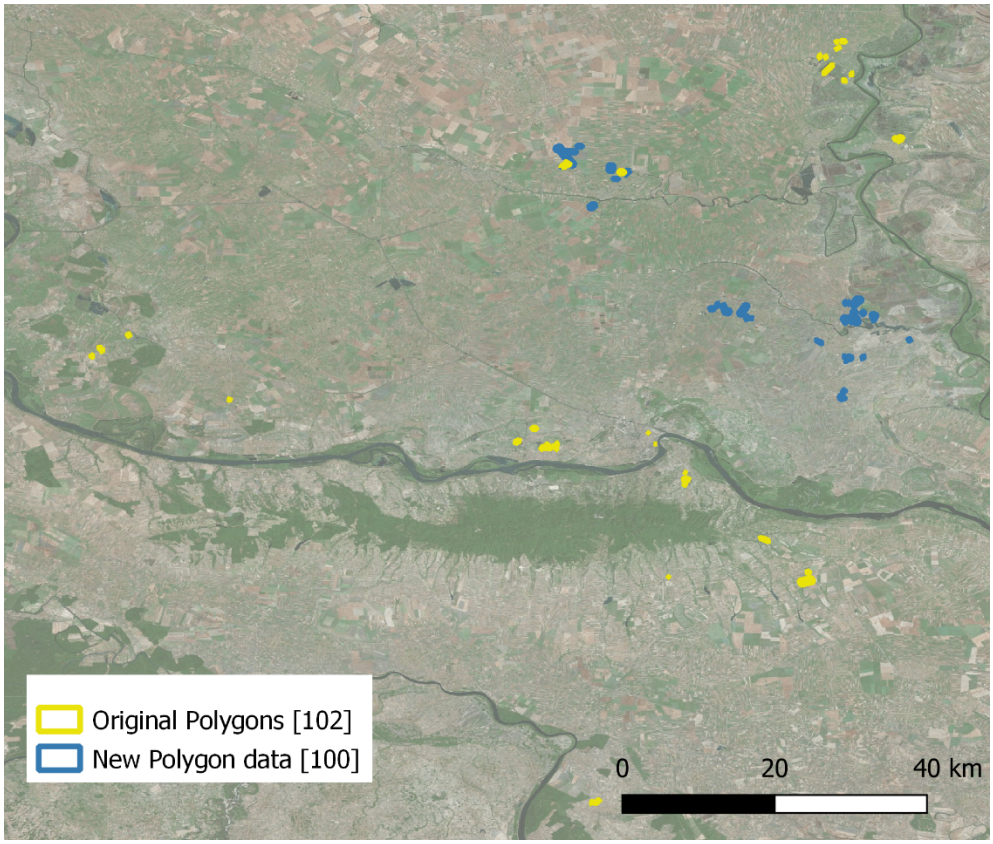

Figure 2: All polygon data collected as part of the CropSupport pilot

The LandSense Polygon Topology check was performed on all new polygons, and the results are shown in Figure 3. As shown in the figure, the number of overlapping polygons detected was quite high (38 out of 100) compared to the original dataset where only nine overlapping polygons were identified. These results were verified, and detailed examples of overlapping polygons are shown in Figure 4.

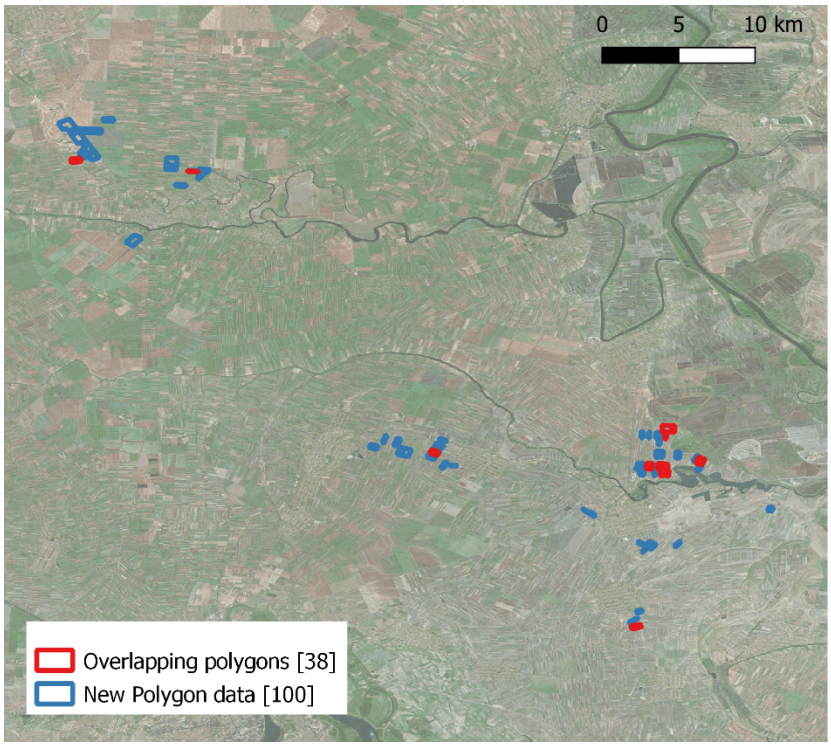

Figure 3: Map of recent polygon data collected with overlapping polygons shown in red

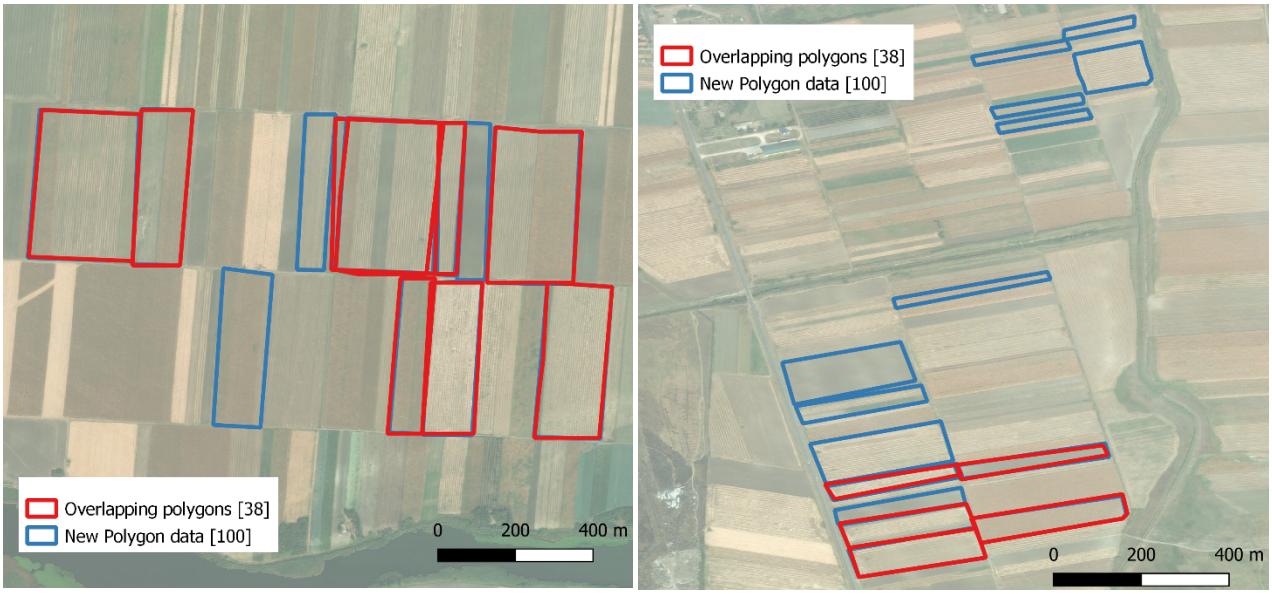

Figure 4: Examples of overlapping polygons

4.2 Photo privacy checks

Photo privacy checks are applicable to all pilots except for the OSMLandUse validation. The results from processing over 1500 photographs collected from the MijnPark.NL, City.Oases, Paysages, CropSupport and Natura Alert pilots are given in this section. Each image was checked for the presence of faces and/or license plates. Where detected, these features were blurred using the algorithms described in D5.2 (Schultz, 2018).

Privacy checks were verified manually for all images, and the resulting statistics for all the image data processed are shown in Figure 5 for face detection and Figure 6 for license plate detection. The statistics shown include:

- The number of images where privacy checks were passed correctly (with the percentage in brackets).
- The number and percentage of commission errors.
- The number and percentage of omission errors.

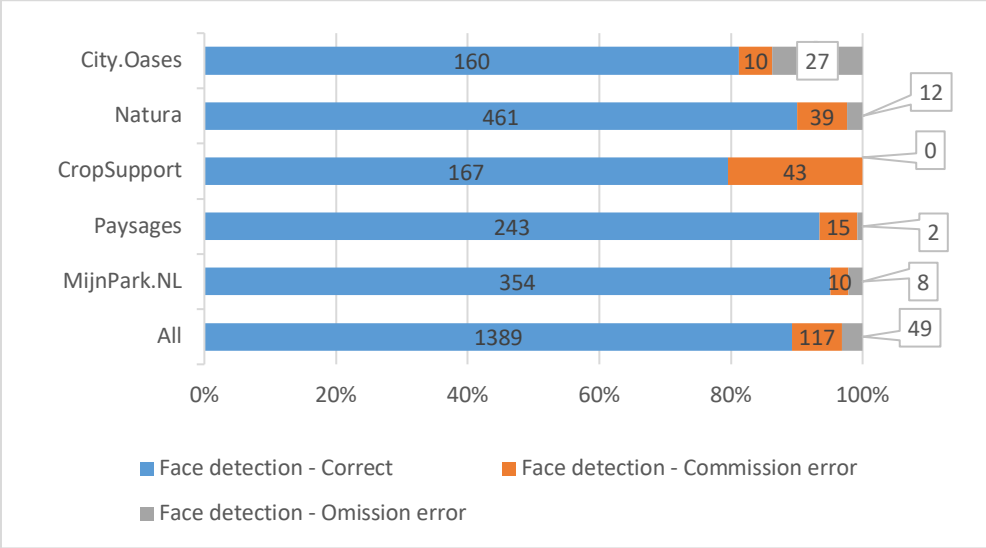

Figure 5: Accuracy results of the face detection service for all pilots

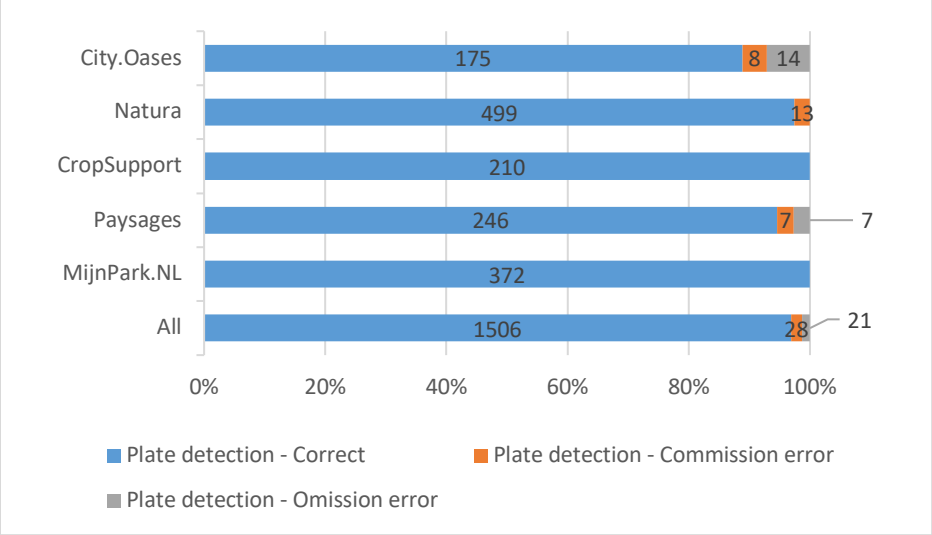

Figure 6: Accuracy results of the license plate detection service for all pilots

The overall accuracy for the face and license plate detection seems to be acceptable, with around 90+% accuracy found for both features. Commission errors are the dominant form of error, with omission errors accounting for a very small percentage (~3% for faces and 1% for license plates). Since the primary function of the privacy check is to ensure that the specific feature is removed/obscured from the data, it could be argued that the overall accuracy is actually over 95% (97% for face detection and 99% for license plate detection) if we focus solely on the more critical error of omission. These results compare favourably with accuracy levels identified in similar studies of face (Bakhtan et al., 2017) and license plate detection (Zhou et al., 2018).

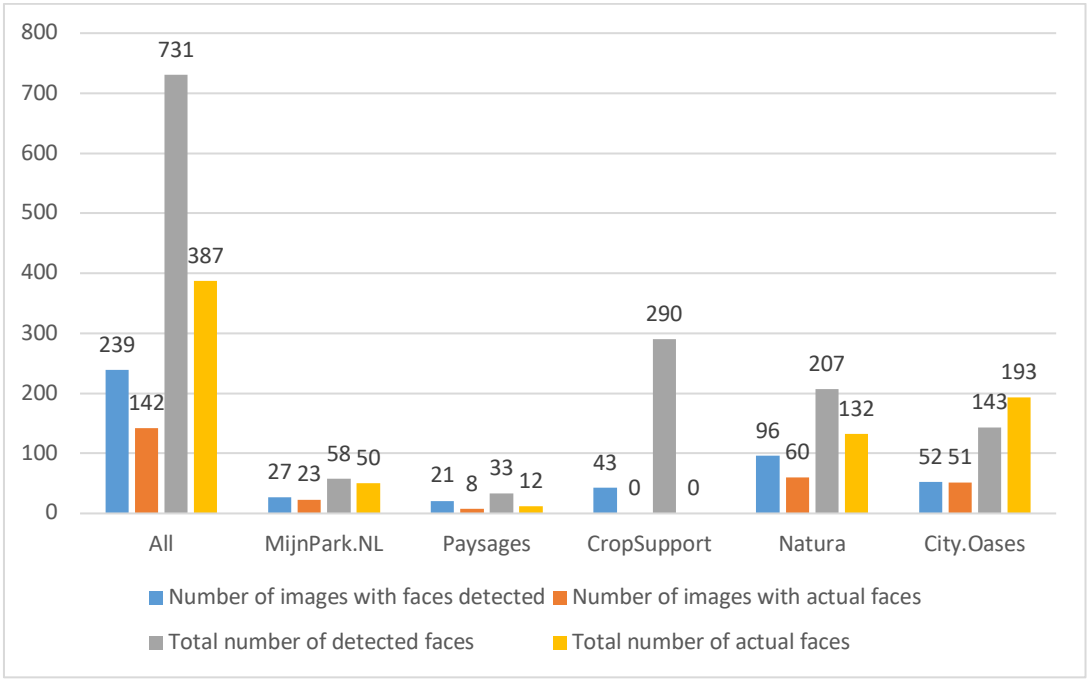

Figure 7: Number of images with faces and the total number of faces (detected and actual) for all pilots

It should be mentioned that, although the overall accuracy was good, the level of commission errors was high, particularly in terms of face detection. The face detection service identified 239 images with features present compared to the actual number of 142 (Figure 7). This is likely to be due, at least partially, to the conscious decision to use low detection thresholds to minimise omission errors, as described in section 3. A large proportion of the face detection commission errors appear to be related to the CropSupport pilot. This is further investigated in the following subsections describing privacy check performance for each individual pilot. These sections will also provide illustrative examples of the algorithm's performance with actual imagery.

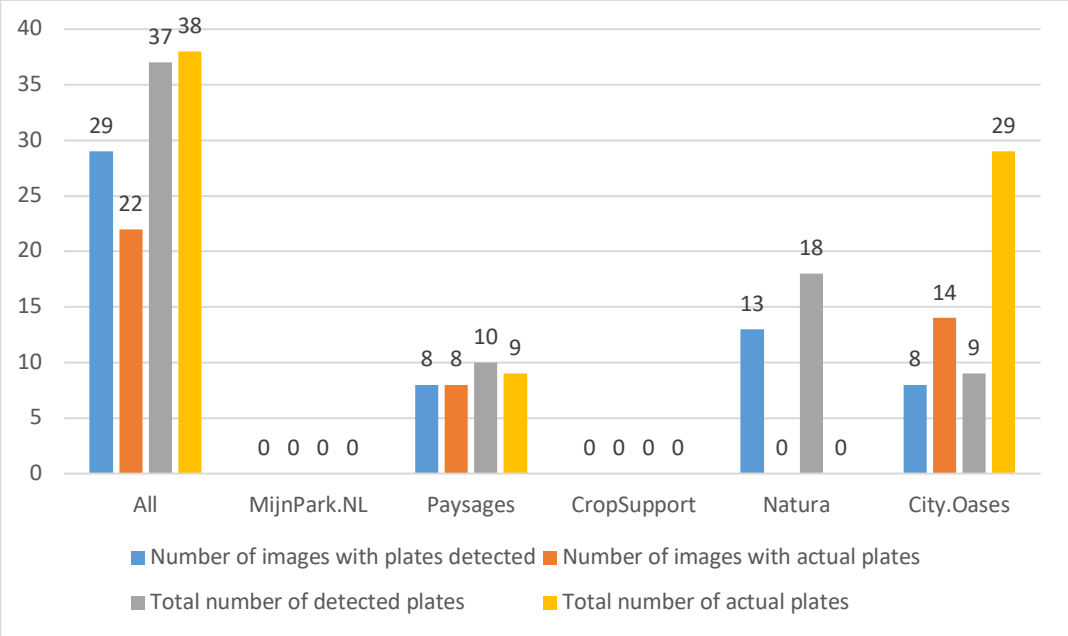

Figure 8: Number of images with license plates and the total number of license plates (detected and actual) for all pilots

4.2.1 City of Amsterdam – MijnPark.NL

Based on the results shown in Figure 5, it is evident that this pilot performs better than average for face detection with higher levels of accuracy than that found for the overall dataset and a greatly reduced level of commission error. The face detector shows good performance in correctly identifying the total number of faces within the images compared to the results for the combined pilot data. An example of this is shown in Figure 9 - the algorithm correctly identified 13 visible faces in the image and blurred their features. Although there are more than 13 people in the image, those in the mid and background do not have clearly visible faces and therefore did not require blurring.

Figure 9: Example image with multiple faces identified and blurred.

From the face detection results, it seems that the algorithm developed is well suited to processing the types of images collected in this pilot (which mainly consisted of images of recreational locations and activities in urban environments).

However, in terms of the more critical omission errors, the face detector did perform slightly worse than the overall average (2.7% compared to 1.6% for the entire dataset). Rechecking the eight omission errors found that it was debatable whether the faces depicted were clearly identifiable This is illustrated in Figure 10, which shows a quarter of the omission errors found in the pilot. In both instances, the faces that are not detected are not fully visible and are borderline errors open to interpretation rather than clear failures to detect a clearly visible and identifiable face.

Figure 10: Two examples of omission errors for face detection - [Left] Image of a person not detected by the algorithm [Right] Full image with multiple people. Manual checks indicate that the algorithm should have detected the face of the person highlighted in red.

No vehicles with clearly visible license plates were present in this pilot's image data. Therefore, this pilot is unable to provide any information regarding the accuracy and/or limitations of the license plate detection tool.

4.2.2 City of Toulouse -Paysages Pilot

In this pilot, faces and license plates were present, allowing for both privacy checks to be fully tested. The overall accuracy is similar for both privacy checks. Errors for license plate detection are split evenly between commission and omissions, whereas for face detection, the majority are commission errors.

Figure 11: Processed image with the face correctly detected and blurred

Figure 11 shows an example of a successful application of the face blurring service. The tool had no problem detecting a face in the foreground even though it was in profile rather than directly facing the camera. Figure 12 shows the two omission errors for face detection identified in this dataset. In both instances, it is notable that the images are quite dark (brightness level 102 and 72, respectively), which is likely to affect the accuracy of the face detection service. In addition, in the case of image 427, the person is in the background of the image and their face blends into the background wall. For image 433, the face is reflected in the wing mirror and is only partially visible, which could also partially explain the identification failure.

Figure 12: The two omission errors for face detection in the Paysages pilot. [Left] image no. 427 and [right] image no. 433

Figure 13 shows an example of a commission error from the face detection service for this pilot. In the processed version of image no 384 (shown on the left), the incorrectly identified face is highlighted in orange. The image on the right shows a zoomed in version of the area where the face was identified. It is apparent that the face detection service mistook a hanging basket of white flowers for a human face. The fact that the original image was quite dark (brightness level 74) meant that the feature in the background of the image, its shape and its colour resulted in confusion, leading to the feature being mistaken for a human face. Increasing the detection threshold would eliminate this type of commission error but could also increase the level of omission errors, which would be a worse outcome.

Figure 13: Examples of face detection commission errors for the Paysages pilot. [Left] Image no. 384, with the error highlighted in orange, [right] zoomed in image of the area highlighted

In terms of the license plate detection service, Figure 14 gives an example of both successful detection and a failure to detect in the same image. The privacy check fails to identify the license plate on the silver car in the left of the image but successfully identifies and blurs the license plate of the vehicle on the right. This may be due to the fact the license plate successfully detected is orientated at a better angle to the camera or it could simply be due to the darkness of the image. However, the fact that the license plate shown in Figure 15 was not identified by the OpenALPR algorithm, despite being more clearly visible, would seem to indicate that the angle of the plate to the camera is the more likely factor.

The performance of the license plate detection service where vehicles actually exist in the images appears to be rather poor. The QA service only successfully identified one out of the eight license plates in this pilot. Due to the small size of the sample, the accuracy, or lack thereof, cannot be completely confirmed. This perceived poor performance is further discussed in Section 5.

Figure 14: Example of license plate detection success and failure in a single image

Figure 15: Example of omission error for license plate detection

4.2.3 Serbia - CropSupport Pilot

As with the MijnPark.NL pilot, there were no vehicles with visible license plates in this dataset. Therefore, no inferences or discussion of the license plate detection tool is applicable to this pilot.

In terms of face detection performance, the accuracy for this pilot is significantly poorer than previous examples, and all errors are commission errors. There are no actual images with faces in the dataset, yet the face detection tool falsely identified 290 faces in 43 of the 210 images. This pilot appears to be the principal source of the majority of commission errors. Examples of some of the images where faces were incorrectly

detected are shown in Figure 16. Based on these results, it would appear that the face detection tool has a problem with this type of image. Images of crops were probably not a part of the algorithm's original training set, which accounts for its poor ability to distinguish between images of crops and faces. Options and alternatives to mitigate this problem will be further explored in section 5 of this report and in the next deliverable D5.7 on good practice guidelines for QA.

Figure 16: Examples of commission errors where crops were incorrectly identified as faces (errors highlighted in red)

4.2.4 Spain and Indonesia - Natura Alert pilot

The Natura alert pilot collected the largest number of images of all the pilots described in this report, accounting for almost a third of the total number of images. Notably, the omission error rate for face detection is higher for this pilot than the examples given so far (although still only 2.3%) and will be further investigated to identify potential sources of error. In terms of license plates, the dataset did contain one image with a license plate present and not detected by the OpenALPR algorithm. On detailed examination, this plate was in the image background, not clearly visible, and was only partially in the image (Figure 17). The use of the OpenALPR algorithm did result in a number of commission errors relating to license plates for this dataset. This was found to be mainly due to the presence of image metadata (i.e., the date and time that the photo was taken along with geotagging information) on the collected images being mistaken for license plate lettering (see Figure 18 for examples).

Figure 17: Single image with partial license plate present and not detected. Two faces also present and successfully detected by the face detection service

Figure 18: Examples of image metadata causing commission errors for license plates in the Natura Alert pilot imagery data. The original image is on the left with the processed image on the right.

Regarding the more critical issue of omission errors for faces, as stated, the number for this pilot is higher than the overall average. Examples of a quarter of the omission errors for this dataset are shown in Figure 19, Figure 20 and Figure 21. In the case of Figure 19, it is clear that shading is obscuring the person's face to

a large degree, and it could be argued that failure to identify the face is, in this instance, acceptable given the darkness of the image.

Figure 19: Example 1 of omission errors from the Natura Alert pilot. There is a person visible in the foreground although the face is shaded by vegetation.

Figure 20 includes three people although only the face of the figure on the right of the building is clearly visible to the camera. Moreover, their features are not clearly identifiable due to their distance from the camera and the brightness of the image. Again, this is a borderline case where it could be argued that the failure to detect the faces is acceptable given that the faces are not clearly recognisable.

Figure 20: Example 2 of omission errors from the Natura Alert pilot – No faces detected

Figure 21 shows an example of where the face detection service did identify a number of faces in the image but failed to detect all of them. Faces of six of the fifteen people in the image were correctly identified and blurred. Of the remaining nine, four of the faces are obscured or facing away from the camera, three are not clearly identifiable, and only the two standing figures on the left and right of image should have been identified, representing a definite failure in detection. It should be mentioned that the image is atypical of the image data collected by LandSense, and it is thought to be a test image taken by users at an event to launch the Natura Alert app.

Figure 21: Example 3 of omission errors from the Natura Alert pilot. Six faces were detected and blurred but a number remain undetected and clearly visible in the image

Privacy check failures, and options for correction and/or mitigation, will be one of the topics further explored in LandSense deliverable D5.7. Despite the errors described, it should be noted that the face detection tool successfully identified most instances of clearly visible faces in the imagery (see Figure 22 for two typical examples).

Figure 22: Examples of successful detection and blurring of faces from the Natura Alert pilot

4.2.5 City of Vienna – City.Oases

The City.Oases app used additional photo functionality that was not present in the pilots discussed previously. Users were able to identify and blur areas of the images using their mobile device's touchscreen. This meant that the majority of the images collected did not require QA checks to be performed. The process of allowing users to identify privacy concerns within the app itself is a topic that will be further explored in the QA good practice guide (Deliverable D5.7).

Of the around 1800 images collected, only 197 images were processed using the QA tools and are included in this report. These images were those identified as having privacy or quality concerns. The filtering was performed manually using Picture Pile (Danylo et al., 2018), which enables users to quickly categorise a set of images. LandSense team members undertook this process as part of ongoing research that will be further described in D5.7. This research is to identify and quantify good practice methods for image processing.

Manually pre-filtering the dataset before processing removed the easier to process images, leaving a dataset of increased difficulty to those described so far in this section. Therefore, the results for this pilot differ from those shown previously, as they represent a dataset of higher difficulty. The rate of omission errors for the images processed was 13.7% (27 images) for face detection and 7.1% (14 images) for license plate detection. In terms of commission errors, for face detection, the error rate was 5% (10 images) and for license plate detection 4% (8 images). Therefore, despite the increased level of difficulty, the QA privacy check tools still managed to achieve a good accuracy rate of over 80% for face detection and 88% for license plate detection.

4.3 Photo Quality checks

Photo quality checks were also applicable to all pilots collecting photographic imagery (all pilots except the OSMLandUse validation). Thresholds to identify blurred and dark images were set at 250 (blur level) and 100 (brightness level) as described in section 3.2. The values for blurring and brightness level were also recorded to allow for post-processing of blurring and brightness using different thresholds if required. Table 2 shows the photo quality results for the image data processed, including a breakdown by individual pilot.

Table 2: Photo quality check results for the image data processed - broken down by pilot

Statistic	All Pilots	MijnPark.NL	Paysages	CropSupport	Natura Alert	City.Oases
Number of blurred images (%)	41 (2.6%)	5 (1.3%)	11 (4.2%)	0 (0%)	0 (0%)	25 (12.7%)
Number of dark images (%)	431 (27.7%)	187 (50.3%)	84 (32.3%)	7 (3.3%)	72 (14.1%)	81 (41%)
Average blurring level	249.2	251.7	252.1	254	253.9	223
Average brightness	111.4	99.3	108.5	122.7	123.0	95.5

CropSupport and Natura Alert have the most images that have passed the two quality checks with zero blurred images and low percentages of dark imagery in comparison to the other pilots. This is potentially due to the nature and context of the image capture between the pilots. Paysages, City.Oases and MijnPark.NL are capturing images in dynamic urban environments at differing times of day whereas CropSupport images mainly consist of static fields of crops taken during daytime conditions in good weather.

4.3.1 Blurring checks

Blurring checks identified 41 blurred images out of the total 1555 images checked. The majority of these are found in the Paysages and City.Oases pilots, with no blurred images identified in the Natura Alert and CropSupport pilots. This is potentially due to the context in which the images were collected. Many of the blurred images appear to have been taken from a moving vehicle (Paysages), in dynamic urban environments (City.Oases) or in very low light conditions, which will have contributed to their poor quality. The overall low rate of blurring was verified by manual checking. Only three images across all pilots was identified, where the accuracy of the blurring result was found to be questionable. These anomalies will be further investigated in LandSense deliverable D5.7.

The ability of the QA tools to differentiate different degrees of blurriness is illustrated in Figure 23 and Figure 24. The two images shown in Figure 24 fall just below the blurring threshold and are only slightly blurred on close inspection. In comparison, the two images in Figure 23 are heavily blurred on even a cursory viewing.

Figure 23: Example of heavily blurred images (both from the Paysages pilot). Left image blur level 216. Right image blur level 162

Figure 24: Example of slightly blurred images. Left from the MijnPark.NL pilot, blur level 249. Right from the Natura pilot, blur level 248

4.3.2 Illumination checking

Illumination checks identified 431 out of the 1555 images as falling below the brightness threshold used. The illumination-checking tool appears to perform as expected with higher brightness scores found for brighter images and lower scores for the darker images. To illustrate this, examples of images with both very high brightness levels (Figure 25) and very low brightness levels (Figure 26) are provided.

Figure 25: The two images with the highest brightness level scores – Both are from the Natura Alert pilot. [Left]Brightness level 187, [right] Brightness level 217.

Figure 26: Two images with very low brightness level scores. The image on the left is from the MijnPark.NL pilot and has a brightness level of 47. The image on the right is from the Paysages pilot and has a brightness level of 46.

The illumination check provides an indication of the overall brightness of an image but there are limitations to the approach used. As described in section 3, the method employed calculates a single brightness value for the whole image based on the average brightness level of each pixel in the image. This can lead to results

where the overall brightness can be misleading. For example, details in the foreground could be in heavy shadow, making them difficult to distinguish, but extreme brightness in the background of the image (e.g., due to a sunny sky) might lead to an overall brightness level that passes the brightness threshold chosen.

As hypothesised in Section 4.2, there appears to be a link between the brightness of an image and errors in the privacy checking, with darker images more likely to trigger an error. Due to this, examples of images with low brightness levels can already be seen in Figure 12, Figure 14, Figure 15 and Figure 19. The actual brightness levels for these images are as follows: Figure 12 [left] (brightness = 102), Figure 12 [right] (brightness = 72), Figure 14 (brightness = 106), Figure 15 (brightness = 72) and Figure 19 (brightness = 68). Further discussion of the links between image quality and privacy check accuracy is given in section 5.3.

4.4 Positional accuracy

Data on the accuracy of the GPS were provided by the City.Oases, MijnPark.NL and CropSupport pilots. Figure 27 shows the distribution of the positional accuracy for the three pilots. The distribution for the two urban-based pilots (MijnPark.NL and City.Oases) is bimodal, showing two distinct clusters of accuracy at approximately 4m and 10m. The rural CropSupport pilot displays a Poisson distribution, also peaking at around the 4m point. It is hypothesised that the secondary peak for the urban pilots (at around 10m accuracy) is likely due to obstructions in the urban environment, which reduced the GPS accuracy. This is further examined in the relevant subsections for each pilot. The mean accuracy for all pilots was similar, with averages ranging from 6-7m. The bulk of the GPS data collected suggests that is fit for further analysis, since the data can be distinctively linked to observed features on the ground.

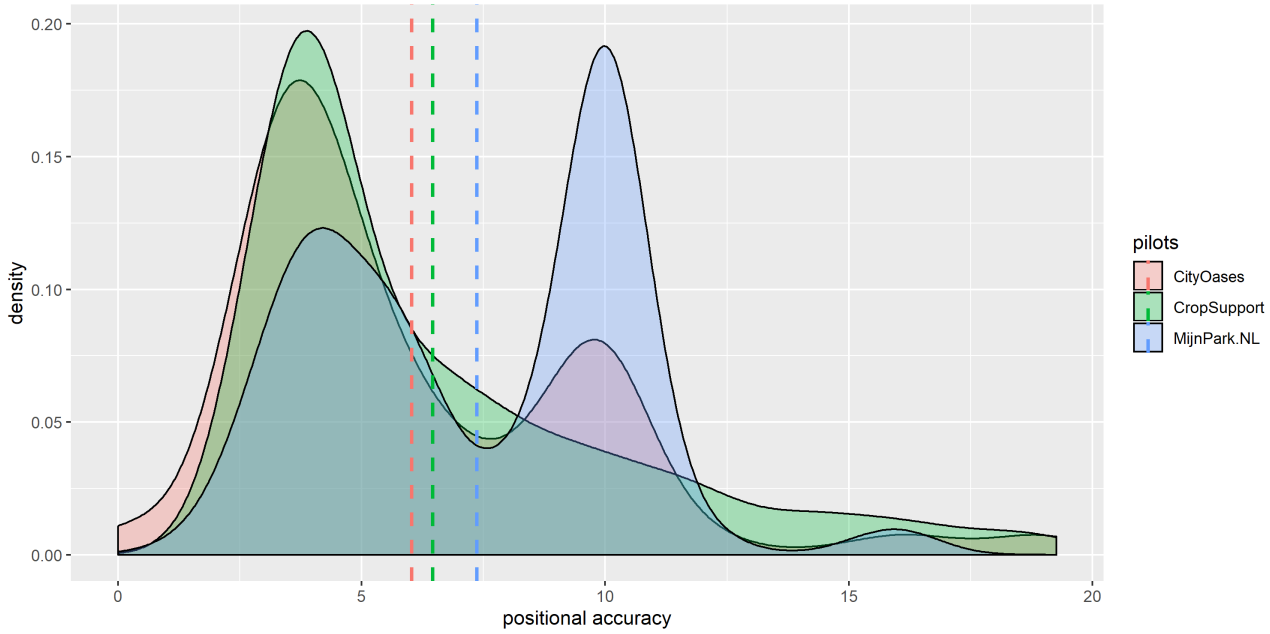

Figure 27: The distribution of positional accuracy results (in metres) by pilot

Figure 28 shows a box and whisker plot for positional accuracy results for each of the three pilots. As can be seen, the medians for the CropSupport and City.Oases pilots are almost identical, with the MijnPark.NL pilot

showing a lower median accuracy and larger interquartile range. However, it is evident that the majority of the positional accuracy results for all pilots fall within the 4-10m range. It is of interest that the rurally based CropSupport pilot shows a large number of outliers. This will be further discussed in section 4.4.3.

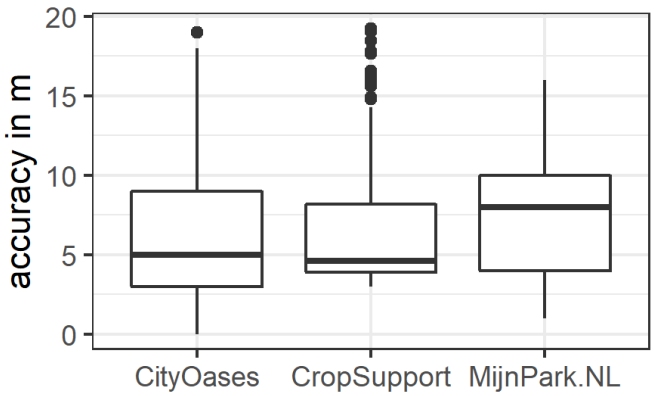

Figure 28: Box and whisker plot of positional accuracy results by pilot. The box represents the interquartile range and the line in the box shows the median. This format is used for subsequent box and whisker diagrams.

Table 3 gives a statistical summary of the results for each of the pilots and includes what is referred to here as the inclusion rate. As described in Section 3.3, any point with positional accuracy >20m was excluded from the analysis. The inclusion rate column shows the number of points that were included in the analysis and the percentage of the total number of points (with accuracy data) this represented. For example, the CropSupport pilot had 207 out of 211 data points with accuracy levels within the 20m range, representing 98% of the data collected. The low inclusion rate (0.82) of City.Oases was due to large outliers likely caused by the large heterogeneity of collection locations across Vienna.

Table 3 : Statistical summary of positional accuracy results

Pilot	max	min	median	mean	Inclusion rate
MijnPark.NL	16	1	8	7.36	361/377 = 0.96
CropSupport	19.26	3	4.6	6.46	207/211 = 0.98
City.Oases	19	0	5	6.02	443/543[3] = 0.82

A detailed breakdown of results for each individual pilot is provided below.

4.4.1 City of Amsterdam – MijnPark.NL pilot

As shown in Table 3, the results for this pilot show the lowest overall accuracy, with the mean and median at 7.36m and 8m respectively. The bimodal distribution of the accuracy results also shows a much higher

[3] Note that only 543 out of the 878 user observations are included here, as 335 data points did not include GPS accuracy data.

secondary peak around the 10m point in comparison to the City.Oases distribution (Figure 27). The reasons for the reduced accuracy in this case are assumed to be related to the local environmental conditions (e.g., the presence of obstructions to a clear GPS signal) likely caused by trees or proximity to buildings.

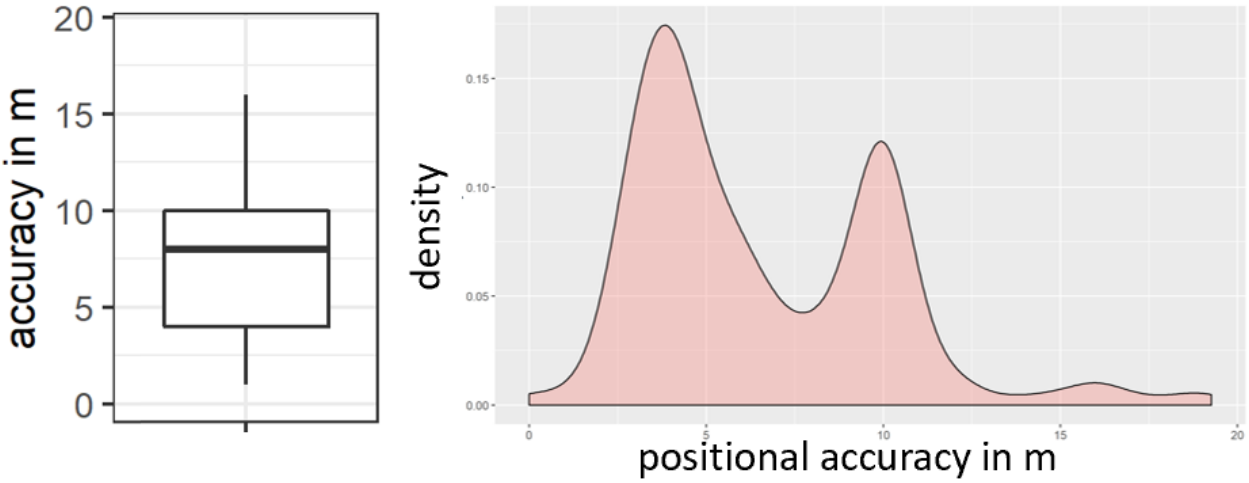

Figure 29: Positional accuracy for the MijnPark.NL pilot data

Figure 30 shows the spatial distribution of the accuracy for each point. The positional accuracy data were categorised into four bands: high (0-5m), moderate (5-10m), low (10-20m) accuracy, and a final band for those values exceeding the 20m threshold used. Observations showing lower levels of positional accuracy were also found to be close to tall buildings and near trees. Observations with higher levels of accuracy were taken in areas with a clear view of the sky.

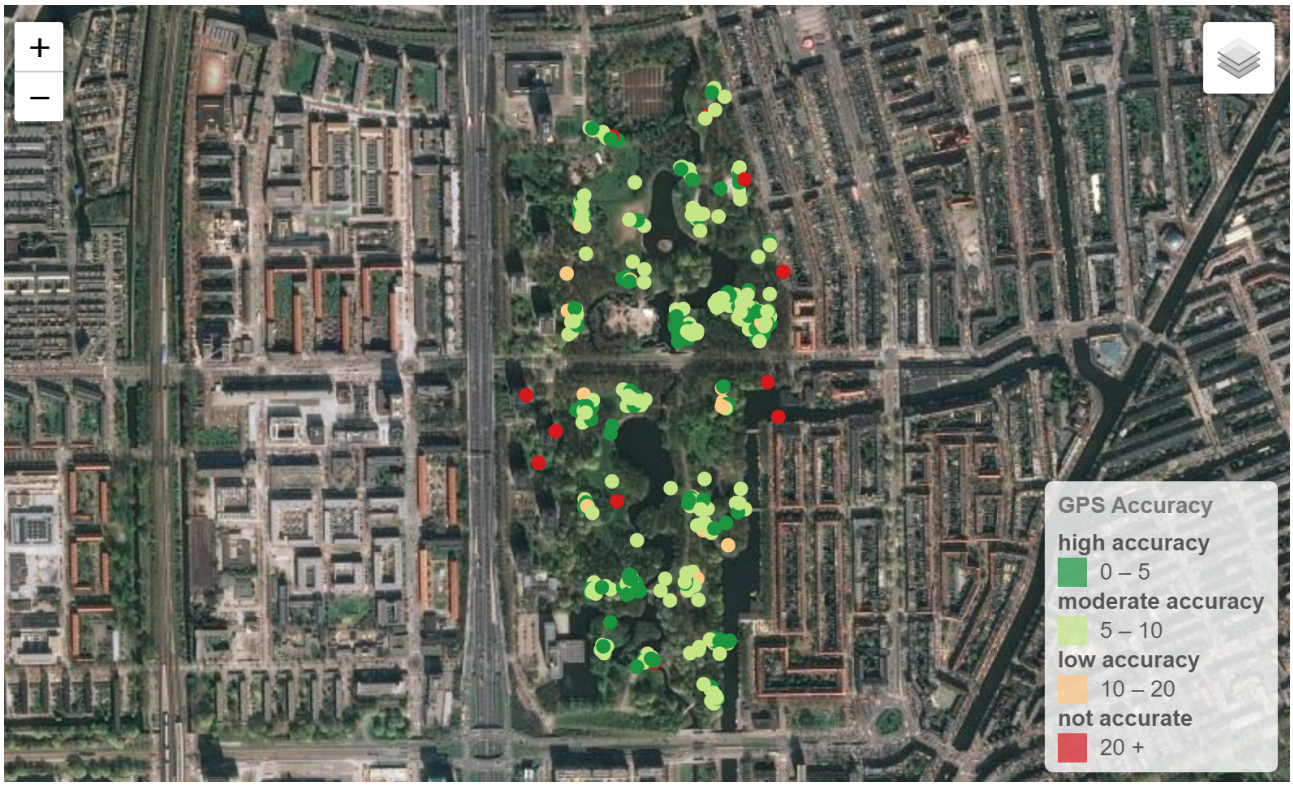

Figure 30: Positional accuracy results by observation point for MijnPark.NL pilot data

4.4.2 City of Vienna – City.Oases pilot

The positional accuracy results for this pilot are also bimodal but with a much smaller secondary peak at the 10m accuracy level (Figure 31). It also has the highest mean positional accuracy (6.02m) but a high dropout rate, where 82% of the points were below the 20m threshold. As a result, about a fifth of all observations were not usable for further exploitation. This pilot would benefit from improved GPS accuracy and/or advice for contributors to give the GPS additional time to increase the accuracy.

As in the previous example, the positional accuracy data were categorised into four bands and mapped (Figure 32). As found with the MijnPark.NL pilot, lower positional accuracy was correlated with proximity to potential GPS obstructions. In contrast to the CropSupport and MijnPark.NL pilots, the observations in this pilot were captured within a highly varying setting of potential phenomena obstructing GPS accuracy including varying weather conditions, presence of large buildings and other potential obstacles.

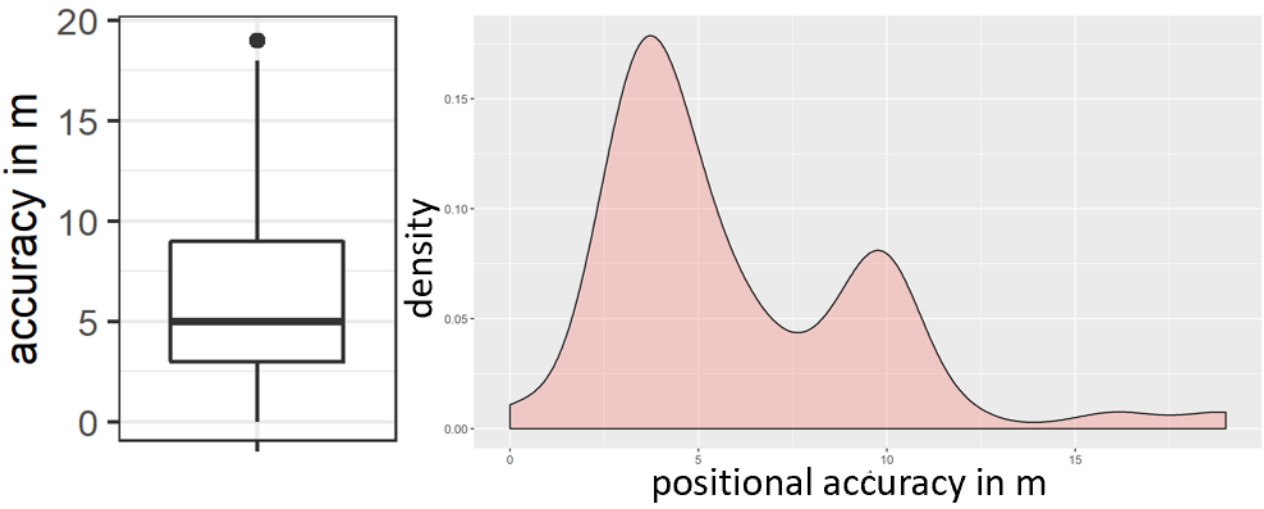

Figure 31: Positional accuracy results for the City.Oases pilot data

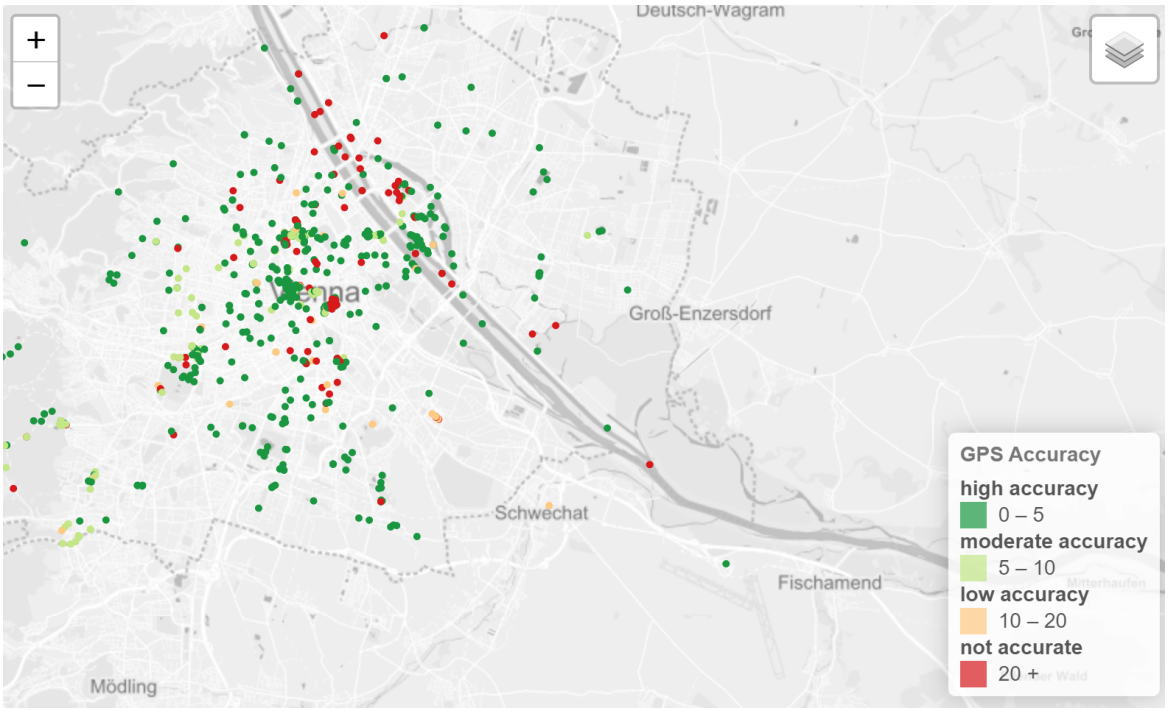

Figure 32: Positional accuracy by observation for the City.Oases pilot data

4.4.3 Serbia – CropSupport pilot

The CropSupport pilot data were collected in rural locations. These were generally in wide-open spaces near agricultural sites. This led to a distribution of positional accuracy results that differs from the bimodal distributions seen for the urban-based pilots (Figure 33). The median and mean values of positional accuracy are very similar to those found for the City.Oases pilot (around 5m and 6m, respectively). There were a number of outliers (as shown in the box plot in Figure 33). The reason for these is not fully known, but it may

be due to issues such as the accuracy of the GPS chip in the mobile devices used to log the observations or it could be related to the meteorological conditions when the observations were made.

The mapping of the positional accuracy results is shown in Figure 34. As can be seen, the majority of observations show high levels of positional accuracy. There are almost no examples of inaccurate positional accuracy displayed. There are a few examples of lower accuracy that appear to be clustered in the NW and SE of the map.

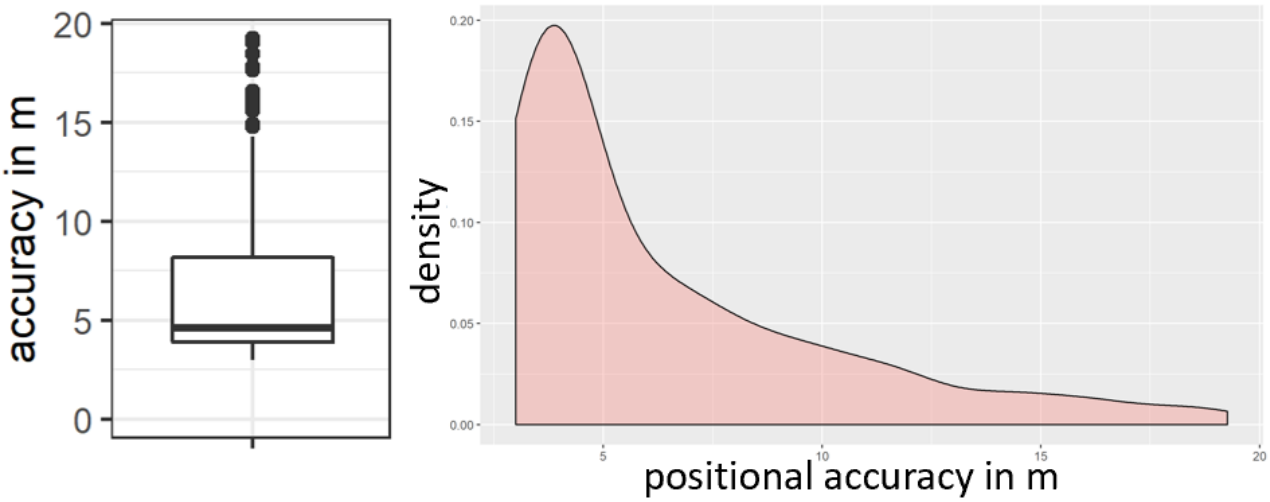

Figure 33: Positional accuracy results for the CropSupport pilot data

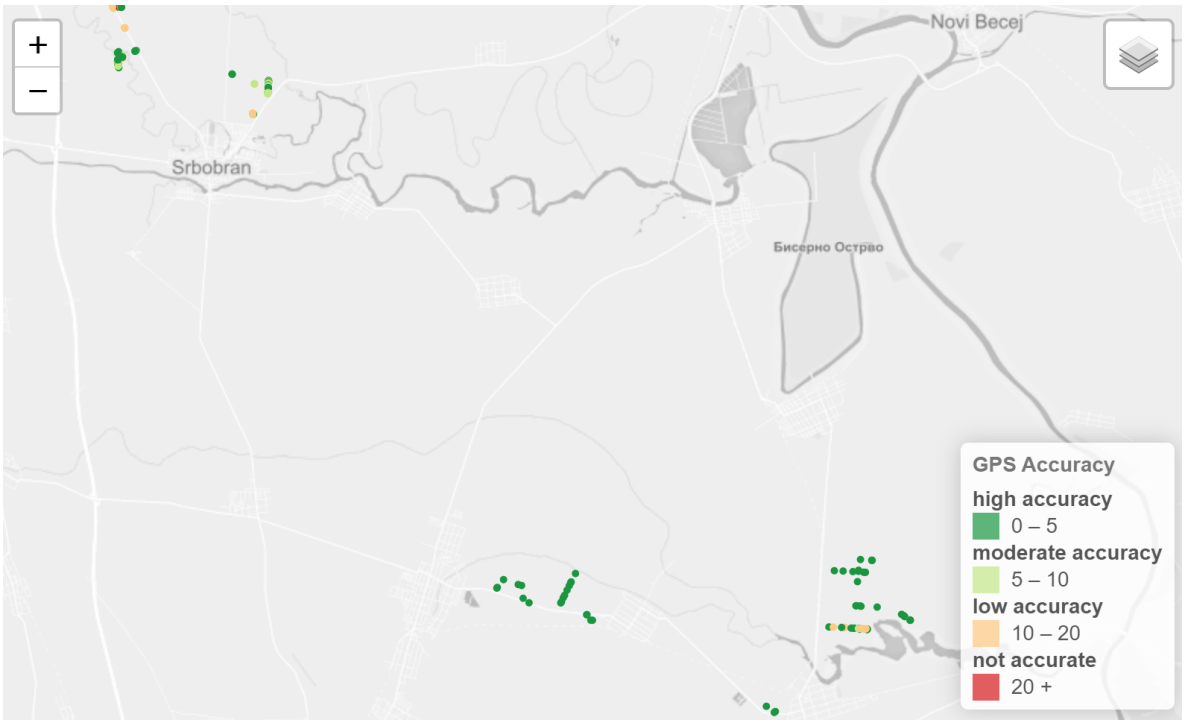

Figure 34: Positional accuracy by observation for the CropSupport pilot data

4.5 Positional offset

Positional offset captures tolerance distances among captured data and their relation to guided points or polygons (D5.2). Positional offset is integrated within the apps and is applied during data capture or production, so no additional post-processing on the LEP is required. An earlier version of the City.Oases app provided a grid of guided points, where a contributor was required to perform data capture within at least a 30m proximity to the guided location. As the app evolved, this feature was removed, allowing contributors to collect observations freely at any given location. MijnPark.NL uses guided locations with observations clustered around those points where a proximity constraint of less than 30m was implemented within the app. Ground observations captured in CropSupport must be linked to polygons drawn in a web form, where both objects must be within at least a 25m proximity of one another in order for a successful link to be made. Since the positional offset is applied during data capture as stated above, there is no resultant data for analysis in this section.

4.6 Contributor Agreement

Contributor agreement was produced for data points of the same feature collected by two independent sources. If the two sources agree, a high contributor agreement was produced, while the opposite is true for low agreement. Contributor agreement was first introduced for the EuroGEOSS 2018 mapathon in Geneva for OSMlanduse validation to identify reliable reference data collected by citizens. Subsequently the method was applied to Paysages, City.Oases and MijnPark.NL. However, as the pilot evolved, the calculation of contributor agreement was dropped for City.Oases as this was no longer relevant. Figure 35 illustrates the contributor agreement density for the LandSense urban pilots. The density distribution of the contributor agreement depicts similarity among the OSMlanduse and Paysages pilots with higher agreement for OSMlanduse. Although both pilots focused on the interpretation of land use, the Paysages pilot was more complex, asking contributors to evaluate changes rather than the status of land use (OSMlanduse validation). This, in turn, could explain the variations in agreement although both were performed using LACO-Wiki (D5.4). Within the Mijnpark.NL pilot, the distribution is bimodal, with both modes less than 0.5. However, for this pilot, agreement on emotions was compared and will be explained in greater detail within the respective

section.

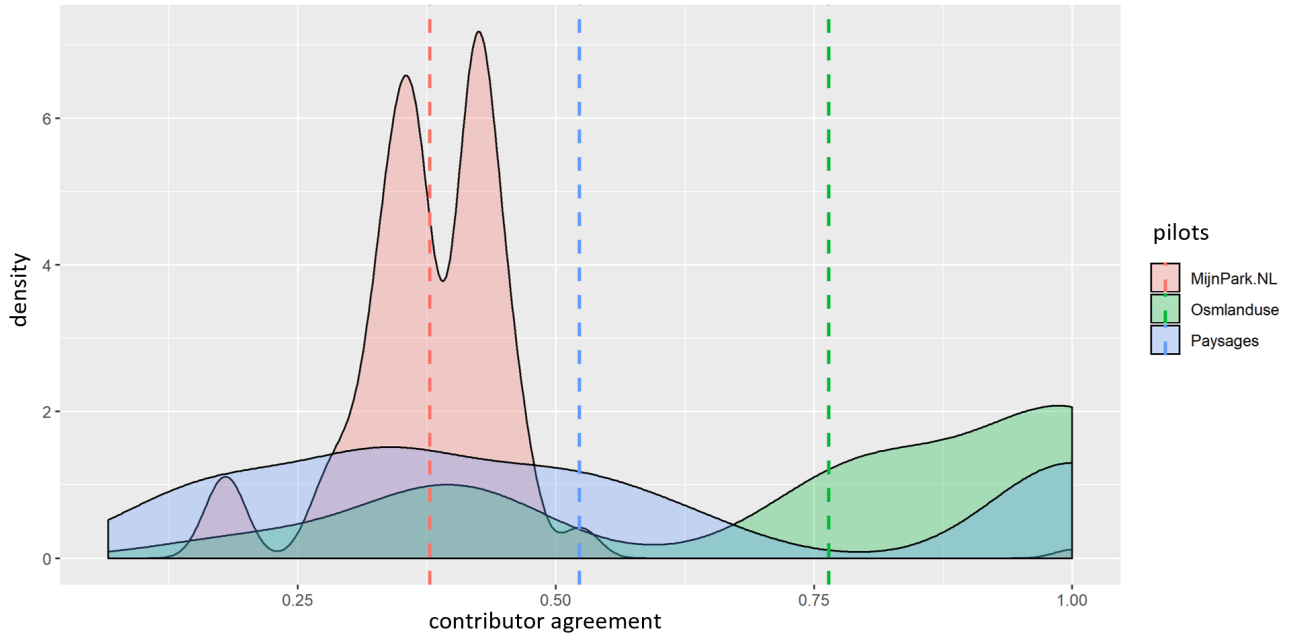

Figure 35. Contributor agreement for the pilots: Amsterdam (MijnPark.NL), UHEI (OSMlanduse), and Toulouse (Paysages), 0 (complete disagreement) to 1 (perfect agreement).

Figure 36 provides further insights, revealing that the OSMlanduse validation pilot has the highest median (0.8) followed by Paysages (0.5) and Mijnpark.NL (0.38). Most outliers were found for MijnPark.NL and the contributor agreement varies across a small range from 0.26 – 0.52. About 50% of the data points are contained within the narrow range of 0.36 – 0.42, revealing that most contributors have a distinctively different perception about the features observed. Contributor agreement data for Paysages span across most of the spectrum, with more than 75% of the data above a contributor agreement 0.28. The median value suggests that contributors neither strongly agree nor disagree; hence, the contributor agreement may provide a useful filter for splitting observations into groups of high and low agreement, although many points may then be considered unfit for use. OSMlanduse observations were the most homogenous and categorically checked by exactly 5 contributors, as the related mapathon was highly guided with simple choices and frequent instructions. More than 50% of the data collected are characterized by high agreement.

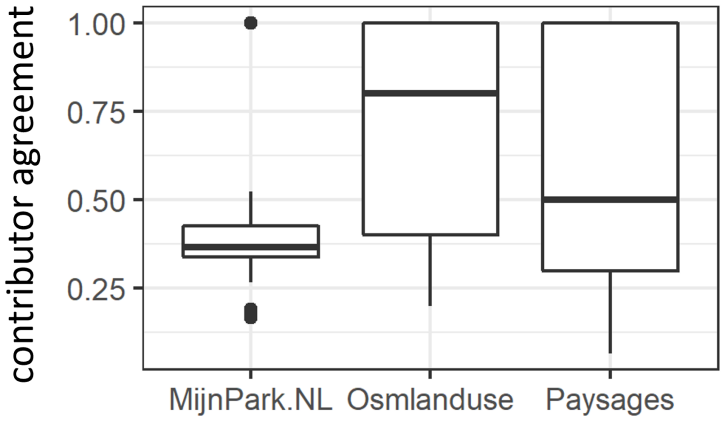

Figure 36. Box plots for contributor agreement for the pilots: Amsterdam (MijnPark.NL), UHEI (OSMlanduse), and Toulouse (Paysages)

Table 4 shows that the range of contributor agreement was the highest for Paysages, followed by MijnPark.NL and OSMlanduse. The table also provides statistics on the contributors per visited point. For Paysages, the largest crowd was mobilized, where contributions for each point varied from 3 to 6. The second largest crowd was mobilized by Mijnpark.NL with 91 contributors, a mean of 12.6 contributions per point and a range of 2 to 40 contributions. OSMlanduse was characterized by few contributors but there were always 5 contributors for each point available, thus producing a very homogeneous data set.

Table 4: Statistics of the contributor agreement for the pilots: Amsterdam (MijnPark.NL), UHEI (OSMlanduse), and Toulouse (Paysages)

		MijnPark.NL	OSMlanduse	Paysages
Contributor agreement	Mean	0.38	0.76	0.52
	Range	[0.17 – 1]	[0.2 – 1]	[0.07 – 1]
Contributors per visited point	Mean	12.6	5	4.27
	Range	From 2 to 40	5	From 3 to 6
	Total number of contributors	91	20	131

4.6.1 City of Heidelberg - OSMlanduse validation

The contributor agreement was calculated only for the data collected in the mapathon in Geneva. Figure 37 shows there is no specific spatial trend evident. Future analyses may focus on the level of contributor agreement for different land use features. During the guided mapathon event, it was observed that the urban fabric and forest classes were easily identifiable for participants, but that they struggled to distinguish between artificial managed green areas and shrubs. The mapathon used a fixed structure, suggesting that high levels of contributor agreement can be achieved by increased participant guidance and good design.

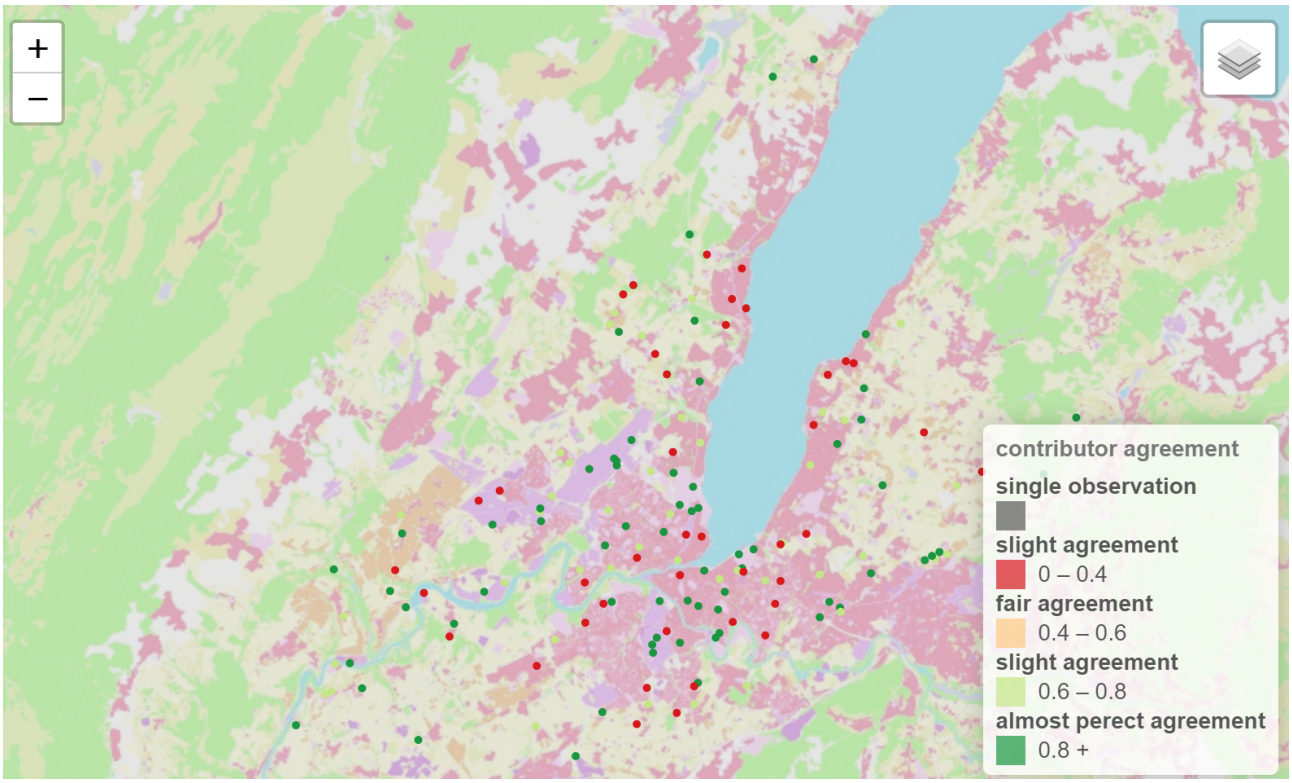

Figure 37: The spatial distribution of contributor agreement for the OSMlanduse pilot in Geneva

4.6.2 City of Toulouse – Paysages pilot

The map in Figure 38 shows the spatial distribution of the coefficient of agreement values for each site. No global trend emerges from this visualization. Moreover, the use of the Local Moran's I statistic indicates that no spatial autocorrelation is present except for a small area in the northern part of the study area.

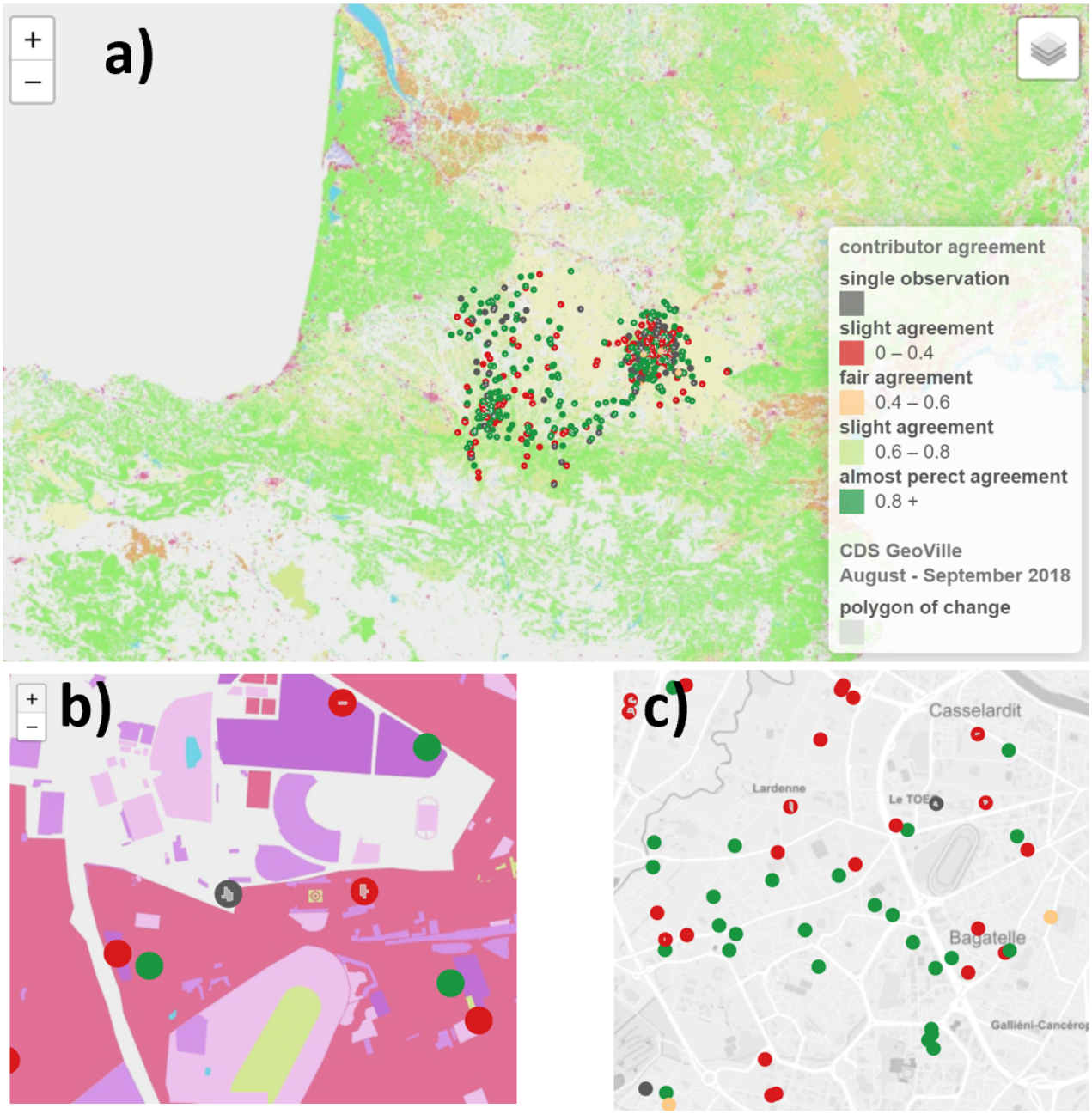

Figure 38: The spatial distribution of contributor agreement for the Paysages pilot

For the Paysages pilot, 650 changes were detected by the LandSense CDS. These changes were validated by different contributors. In total, 2778 ground reference data labels were collected. Each change was labelled by between 3 and 6 contributors in a series of mapathons, with most sites labelled by 4 contributors (Figure 39).

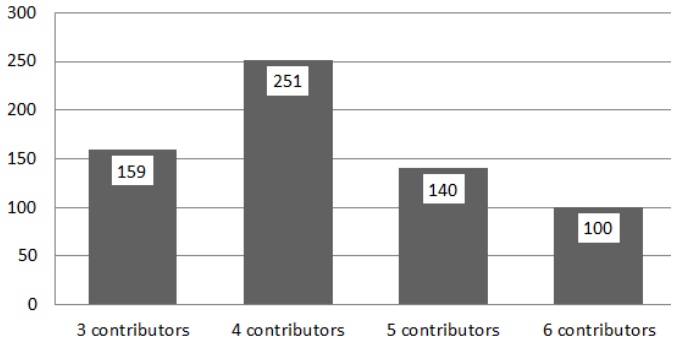

Figure 39: The number of times a site was labelled by 3 to 6 contributors

All levels of agreement observed are depicted in Figure 40. We can observe that when six contributors labelled the same change, they are in perfect agreement very few times (13 labelled sites) and that the total disagreement (contributor agreement = 0) occurs mainly where sites are labelled by three contributors. In contrast, contributor agreement is high (equal to 1) where three or five contributors label the same site. Contributor agreement equals 0.5 and 0.6 for four and five contributors, respectively. Based on this analysis, we can recommend that the total number of contributors labelling the same site should be between three and five.

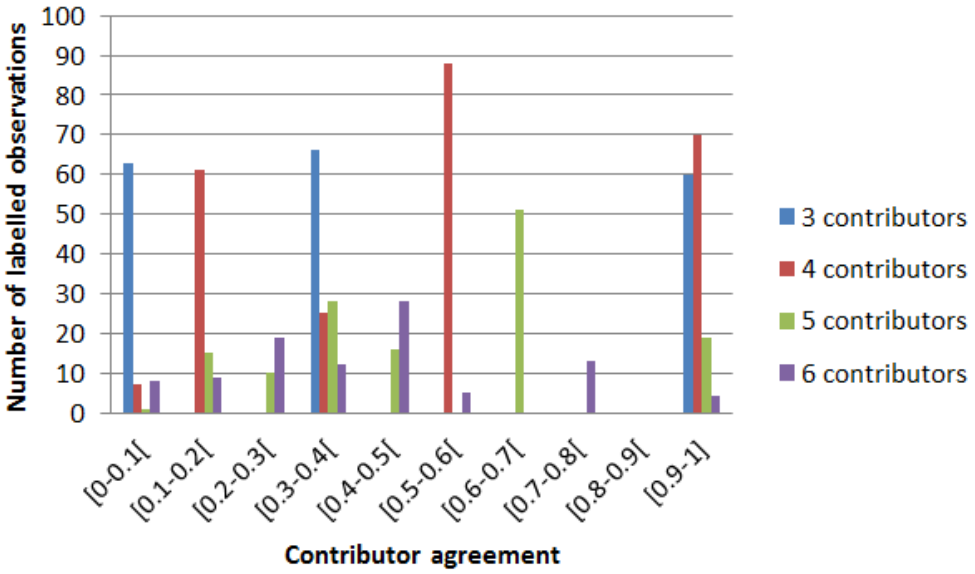

Figure 40: The frequency of the degree of agreement for all changes

4.6.3 City of Amsterdam - MijnPark.NL pilot

Contributor agreement for this pilot was experimental, testing to which extent subjective and emotional perceptions might be suited for its application. Figure 41 shows the agreement in terms of satisfaction with the trees, but the dataset does not reveal if contributors generally like the trees or are disappointed by them. Perception regarding the satisfaction of available benches and their quality is very low, suggesting that

contributors disagree regarding their satisfaction with the available benches. A similar pattern is present regarding the agreement of contributors for their satisfaction with paths.

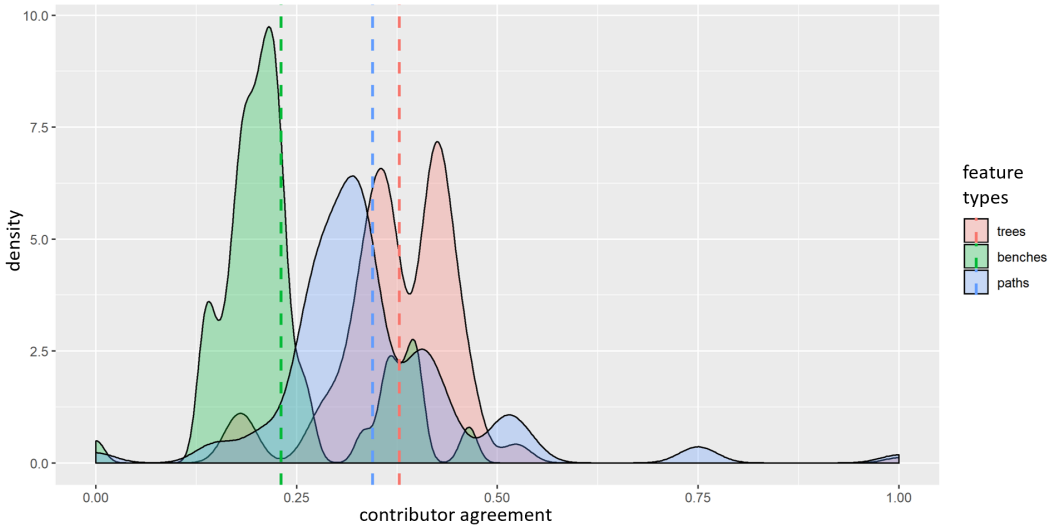

Figure 41: Contributor agreement for MijnPark.NL regarding satisfaction with trees, benches and paths 0 (complete disagreement) – 1 (complete agreement)

Figure 42 shows that the satisfaction with paths is characterized by slight agreement for almost the entire site (Amsterdam Rembrandt Park) with the exception of a small area in the northeast. For benches, the pattern is similar, but with more clusters of fair agreement and one cluster of almost perfect agreement where the respective site on the east side of the park is a playground including benches. Contributor agreement for satisfaction with trees varies by site, ranging from slight to moderate agreement.

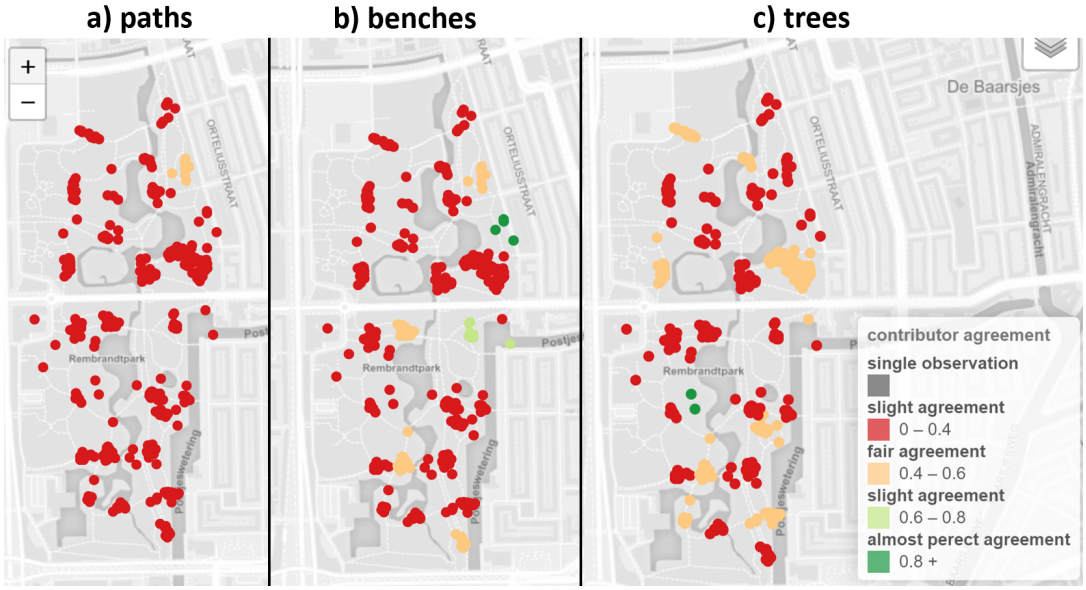

Figure 42: The spatial pattern for contributor agreement for MijnPark.NL regarding satisfaction with trees, benches and paths

4.7 Categorical accuracy

Categorical accuracy was determined primarily for the OSMlanduse validation. Additionally, Paysages made use of this QA service as well as earlier versions of the Natura.alert.app. Either purely thematic or temporal/thematic information was assessed. The OSMlanduse pilot focused exclusively on thematic information regarding the current state of land use while the Paysages pilot focused on the validation of changes produced by the CDS. Deforestation alerts used in the Indonesia pilot in the Natura.Alert.App of BLI were initially produced and validated by UHEI, although at the start of 2020, new deforestation alerts for the app were produced by WUR. Subsequently, the latest developments are outlined, and linkages to other deliverables are provided.

4.7.1 City of Heidelberg - OSMlanduse pilot

The latest results for this pilot were provided in D4.5. However, they are mentioned here briefly for context. In addition to a regular validation mapathon using the LACO-Wiki online land cover validation tool (https://laco-wiki.net), a new area-based categorical accuracy estimation was introduced. Figure 43 shows reference data collected for a site in Heidelberg (Schwetzingen), which was compared with the osmlanduse.org data. OSMlanduse data were also compared to the reference data set produced in Jena, and strengths and weaknesses of OSMlanduse were reported.

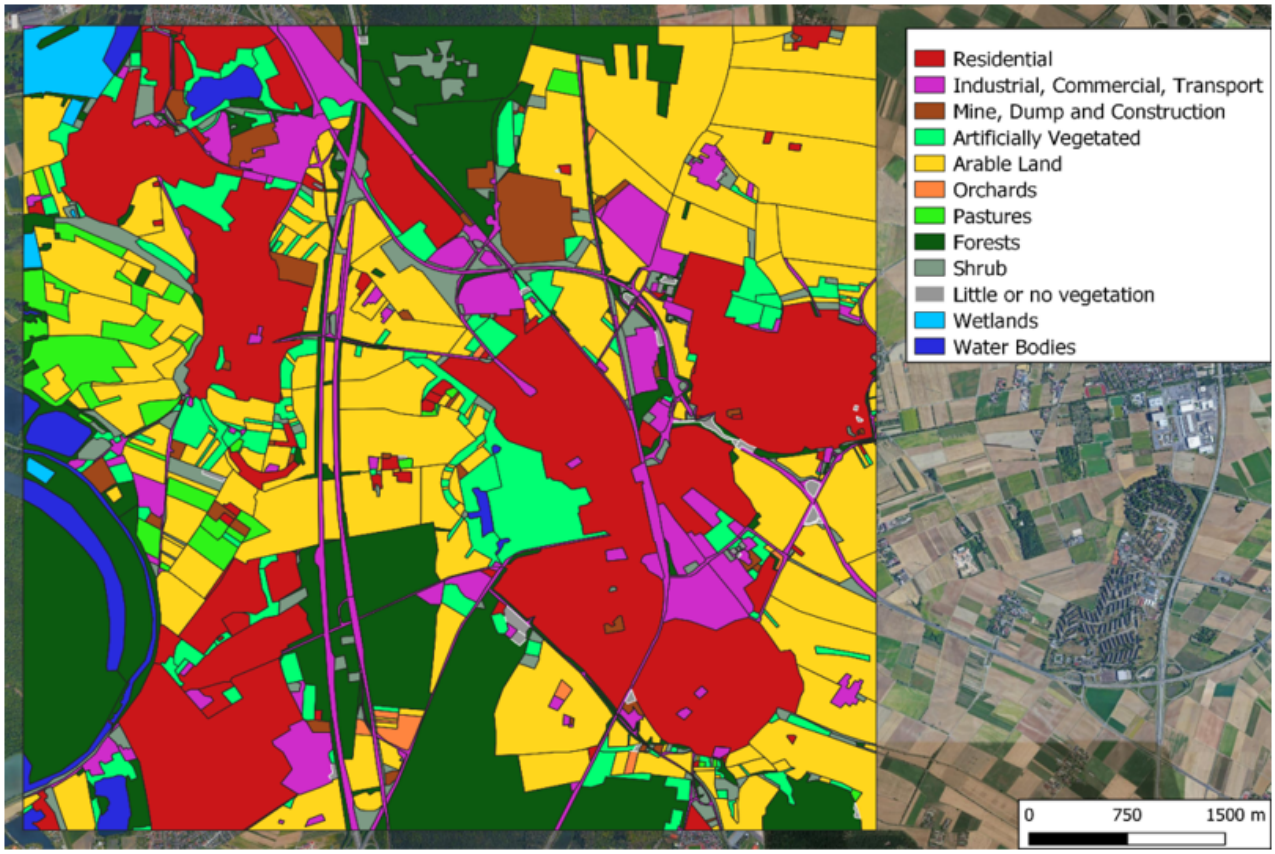

Figure 43: Polygon-based reference dataset produced during the July 2019 DLR/Uni-Jena mapping campaign

Figure 44 suggests a strong similarity between the ground truth (upper left) and the underlying OSM data, whereby sporadic data gaps are visible for the OSM-based classification[4] (upper right). The differences are highlighted between the ground truth data (reference dataset) and the original OSMlanduse data. Grey areas show an agreement between classifications, whereby other colours reveal the respective class assignment for areas of disagreement. The sporadic white patches reveal missing data within OSMlanduse. However, through the validation mapathon, these could be filled in using the online version. Ground truth data and the OSM classification visibly disagree in agricultural areas and for infrastructure. A large part of the data gaps in the OSM classification (upper right) appears to correspond with the class Industrial, commercial and transport units (1.2) in the reference dataset (upper left). Comparing the ground truth and OSM classifications resulted in an agreement of 77.4%. The OSM classification shows a no data proportion of 7.9% for the whole area, while its largest proportion (36.1%) is assigned to Industrial, commercial and transport units (1.2) in the reference dataset. Overall, the best performing classes with over 90% agreement are Urban fabric (1.1), Industrial, commercial and transport units (1.2), Mine, dump and construction sites (1.3) and Forests (3.1). Accuracy is lower than 15% for the classes Pastures (2.3), Shrub and/or herbaceous vegetation associations (3.2), Open spaces with little or no vegetation (3.3) and Inland wetlands (4.1). However, when considering class proportions, less than 1% of the area is classified as Open spaces with little or no vegetation (3.3) or Inland wetlands (4.1) in the reference dataset. The OSM classification shows the highest absolute disagreement for the class Arable land (2.1). Here, an overestimation of 5.8% relates to the class Pastures (2.3) and is clearly visible in the left side of Figure 44. The class Urban fabric (1.1) is also overestimated in the OSM classification and relates to class 1.2 (3.6%) and 1.4 (2.4%). Other areas of disagreement often appear as fragments at the very edges of continuous classifications like rivers, residential areas and streets. Misclassification is common in the context of small-scale structures like standalone buildings inside large continuous areas. If gaps present in the OSM classification are disregarded for the calculation of accuracy measures, the overall accuracy reaches 83% inside the reference area.

[4] (https://osmlanduse.org/#13/8.58427/49.38582/0/)

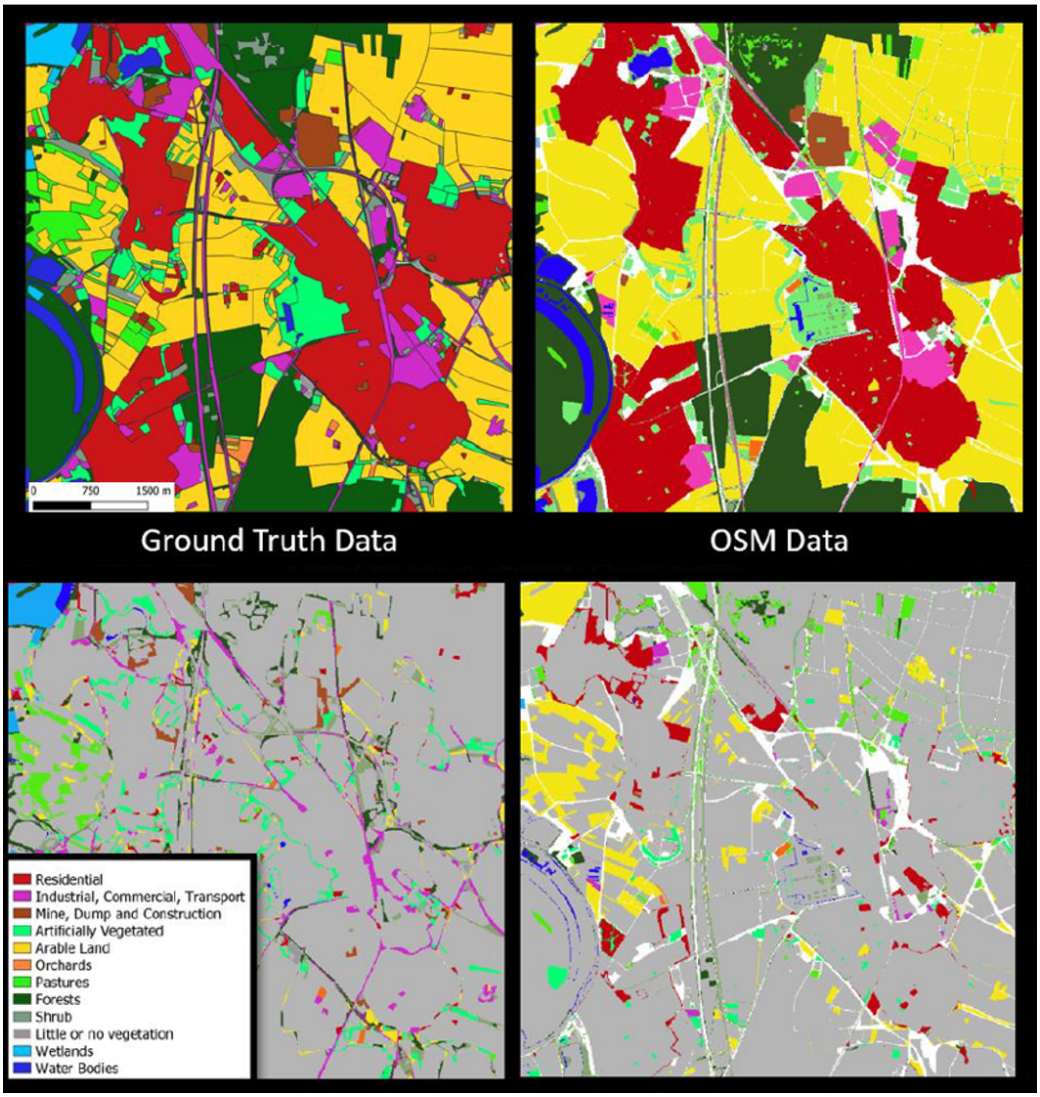

Figure 44: Reference dataset vs OSM data. The top row contains the data sets and the bottom row shows their differences (grey areas depict agreement). The bottom left hand figure shows the reference data collected through the mapathon.

4.7.2 City of Toulouse - Paysages pilot

The outputs from the CDS represent new construction sites (e.g., a building or a road), which were classified into four LU classes: *Residential*, *Industrial*, *Infrastructure*, and *Other*. Since some changes occur over a long period and their final nature may not be apparent until near the end of the process, the classes *Construction in progress* and *Destruction* were defined and added. In addition, since it was possible for a change to be incorrectly classified as having undergone a change (i.e., a false positive), an additional class, *No change*, was defined. Finally, in recognition of the challenges in identifying a change, a site could also be labelled as *Unknown*. In total, therefore, the accuracy assessment focused on an 8-class classification.

The accuracy assessment was undertaken as described in deliverable D5.3, taking the class proportion into account and expressed as the overall accuracy. Labelled changes were removed when the p-values of the agreement coefficient (P_i) were lower than 0.05. If a reference feature had the label Unknown, it was also

removed. Thus, the accuracy assessment was computed for 467 labelled changes (71% of the total number of sites identified by the CDS).

The overall accuracy (as defined in section 3.5) was determined to be 0.81.The user's and producer's accuracies are shown in Figure 45. We can observe that both the UA and PA are greater than 0.85 for the residential class. This means that the CDS performs satisfactorily for detecting residential changes. The UA for the Infrastructure class is low, being equal to 0.5. This is not surprising given the difficulty in detecting Infrastructure usage.

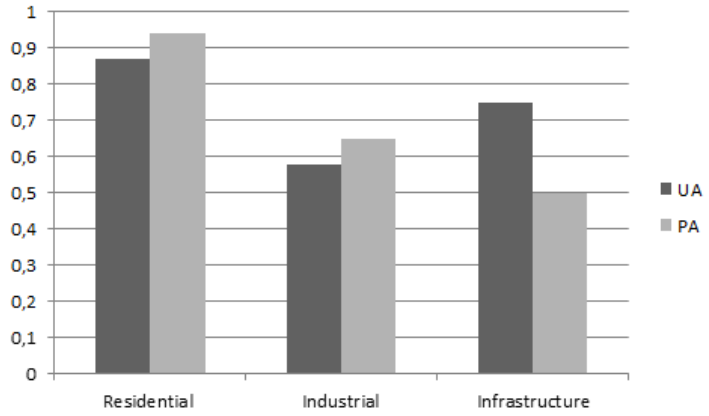

Figure 45. User's and Producer's accuracy for the LU classes from the CDS

If sites classified into *Other* have a p-value of P_i more than 0.05, the reference label is assigned to *No change*. This explains why the UA and PA are equal to 0 and thus, not represented in Figure 45.

Computing the change/no change confusion matrix is difficult since we do not have the reference corresponding to all changes that occurred or locations where changes did not occur. However, as a guide to changes that were not identified, during one mapathon with the experts in LULC, they identified 142 sites of change that were not detected by the CDS algorithm. Nevertheless, as a guide to the accuracy of the classification, only the UA for the change class was computed. The reference change class is defined by considering changes in which the labels were assigned to one of the following: *Residential*, *Industrial*, *Infrastructure*, *Other*, *Construction in progress*, and *Destruction.* For example, if a reference change has the label *Construction in progress*, then the reference change class has the value change. The UA for this change class is estimated to be 0.87.

Figure 46 shows an example where both the CDS and the reference labels agree: Industrial change. The contributor agreement is equal to 0.4.

Figure 46. Example showing a change detected as Industrial (overlaid on an orthophoto for 2016 on the left and 2019 on the right) by the CDS and considered as industrial by the reference.

Figure 47 illustrates the impact of scale on the validation task. The CDS detected a residential construction site and the reference value was labelled as *No change* (with a coefficient of agreement equal to 1 for three contributors). When comparing the two orthophotos from 2016 and 2019, it can be seen that a small change has actually occurred but within an area that was otherwise unchanged. This highlights an uncertainty in the labelling linked to the extent of the area of change.

Figure 47: Example showing a change detected as residential (overlaid on an orthophoto for 2016 on the left and 2019 on the right) by the CDS and considered as *No change* in the reference

4.7.3 Indonesia - Natura Alert pilot

D5.4 section 2.5.3 provided an overview of the categorical accuracy produced for this pilot for the near real-time monitored deforestation alerts. The reader is referred to that deliverable for further information.

5 Discussion

Each of the eight quality assurance tools developed for LandSense has been applied to data acquired during the project. Here, a brief discussion of the results is provided.

5.1 Polygon topology check

Checking for overlapping polygons representing fields of crops in the CropSupport pilot performed as intended. A notable increase in the number of overlapping polygons was identified in the second phase of data collection as described in section 4.1. Methods for reducing polygon topology errors and automating the process of correcting overlaps will be further considered in the good practice guidelines for QA processes to be developed in the next phase of this work package (D5.7).

5.2 Photo privacy checks

The overall results for photo privacy checks given in section 4.2 show a high level of accuracy (90%+), comparable with the accuracy results reported elsewhere in the literature (e.g., Baktan et al., 2017; Mennon & Omman, 2018; Zhou et al., 2018). However, the majority of images collected by the LandSense pilots do not have detectable features, which could inflate the apparent performance of the privacy checks. Figure 48 shows the results of the face detection QA check, where only images with faces present are included. The overall accuracy of the face detection tool is reduced to 65% in this scenario.

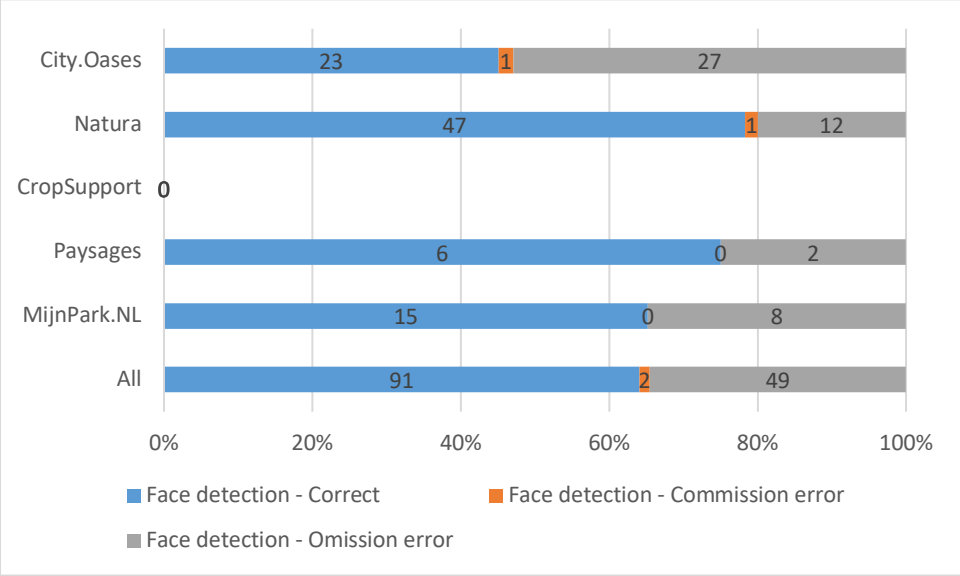

Figure 48: Accuracy of the LandSense face detection checks using only images with detectable features

The reduced accuracy may appear low but is reasonable given the unconstrained nature of the images captured by the LandSense pilots. Unconstrained refers to the fact that faces and/or license plates are captured at various scales, angles and positions within the image and with differing levels of image quality. Yang et al. (2016) describe the difficulties of detecting faces in an unconstrained context, and that even state-

of-the-art face detection algorithms can only achieve accuracy results in the 70% range for these types of images. Hence, the performance of the LandSense face detection tool seems acceptable.

One notable issue with the face detection service was the high level of commission errors in the CropSupport pilot. It was hypothesised that the cause of this error was a lack of images of crops in the training set used to develop the face detector and the high face detection threshold used. The latter was tested by rerunning the face detection test using a lower threshold of 0.5; the results can be seen in Figure 49. It appears that relaxing the threshold is sufficient to eliminate a large proportion of the commission errors for this pilot.

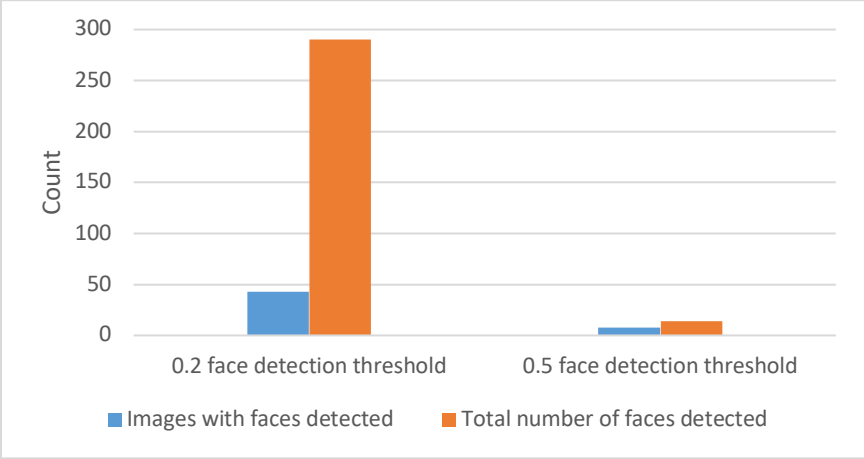

Figure 49: Faces detected using differing levels of face detection threshold for all CropSupport images, with face detection commission errors identified.

In terms of license plate detection, the results based only on assessing images with license plates present are less impressive. Figure 8 shows that there are only 22 images with license plates present. The OpenALPR plate detector does not correctly identify all of the license plates in any of these images; however, it does detect one of the license plates present (see Figure 14). Based on the criteria used for faces, its accuracy is 0%. Even if we calculated the accuracy based on the number of license plates detected, the accuracy is still only 12.5%, neither of which are acceptable levels of accuracy. However, it should be recognised that this poor performance is based on an extremely small sample size compared to the face detection tool.

Most license plate detection systems are capable of achieving performance levels of 90% or higher (e.g., Nejati et al., 2016; Jain et al., 2016). This level of performance is usually for constrained images where license plates are in the foreground of the image, directly facing the camera and clearly visible (Baktan et al., 2017). The performance on unconstrained images is generally greatly reduced. There are implementations of license plate detection systems that focus on unconstrained imagery including variants of the OpenALPR system used for LandSense (Silva & Jung, 2018; Jiao & Fan, 2019).

As part of further work in D5.7 in developing the good practice guide for QA tools, we will investigate potential improvements to the license plate detection services and determine whether these are warranted given that the number of potential license plates present is extremely low (as is the case for the LandSense data analysed here). The potential for retraining the current OpenALPR-based tool will be evaluated along with other options.

5.3 Photo quality checks

Illumination checking has been added to the QA service as described in section 3. Blurring and illumination checks were carried out successfully on the pilot data with no significant errors identified. The two methods used to calculate image brightness and sharpness seem to perform as expected. It was noted that the blurring check score seems to be highly skewed towards the upper end of its range with any minor deviations below the maximum blurring level, leading to significant image blurring. Future work on good practice for QA checks in deliverable D5.7 will investigate other algorithms for testing image sharpness to compare and contrast performance against our implementation.

The illumination check used, based on calculating the relative luminance of the image, is one of many methods that can be used to assess image brightness. As stated in section 3.2, it suffers from the limitation of only providing an overall average brightness for the entire image and cannot detect dark areas within the image. Future work under D5.7 will also investigate and compare other options for assessing image brightness to determine the optimal solution for LandSense and the wider domain of VGI.

It was hypothesised that image quality might be correlated with the likelihood of privacy check errors. Figure 50 shows the distribution of brightness levels by result category for the face detection service for all processed image data. From this, there appears to be some evidence to back up the hypothesis, at least in terms of omission errors. The distribution of omission errors shows a higher frequency of images at lower brightness levels and a small spike at upper brightness levels in comparison to the distribution of images in the other categories. There do not appear to be any significant differences in the distribution between the images with commission errors and those that were correctly processed. It should be noted that the relatively small sample of omission errors does mean that it is not possible to fully confirm the link between image brightness and face detection error.

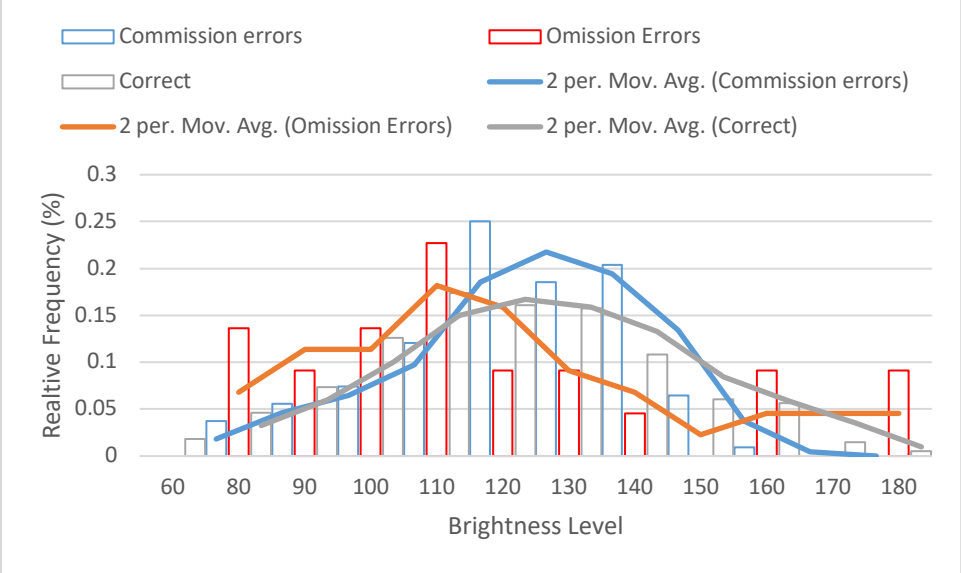

Figure 50: Distribution of face detection results by brightness level and result category

It is not possible to investigate any potential correlation between image blurriness and the error rate for privacy checks, as there were no images classified as blurred with faces and/or license plates present. It is, however, likely that such a link would exist.

5.4 Positional accuracy

The positional accuracy results, as described in section 4.4, showed differences in the accuracy distribution between the urban-based pilots and those observations taken in more open, rural locations. As might be expected, obstructions to GPS signals, e.g., presence of buildings and trees, appear to impact positional accuracy. It was not possible at this stage to analyse other potential sources of positional inaccuracies, such as the GPS accuracy of the mobile devices used. Work to explore other sources will be further investigated in the development of the LandSense QA good practice guide (D5.7).

5.5 Contributor agreement

Contributor agreement was demonstrated across three of the pilot studies and has been shown to be a useful tool in assessing the accuracy of VGI. Contributor agreement may be used as a tool to remove outliers and determine the validity of the observations. Contributor agreement is meaningful if the phenomena under observation objectively possess distinctive unique characteristics such as a certain land use type (*e.g.*, OSMlanduse validation) or the occurrence of a land use change event (*e.g.*, Paysages validation). As emotions vary by individual, contributor agreement is not suitable as a quality estimation tool for such observations (*e.g.*, MijnPark.NL or City.Oases) but can still be of use to determine the levels of agreement beyond a quality perspective for data analysis.

5.6 Categorical accuracy

The previous deliverable D5.5 (Liu et al., 2019) identified a workflow for managing and integrating citizen observations into authoritative mapping processes. The results shown in section 4.7, particularly relating to the OSMLandUse and Paysages pilots, highlighted the great potential of utilising citizen-sourced observations to support and enhance LULC mapping. It was found that specific LULC categories are more suited to the use of citizen-sourced observations, and common types of error were highlighted.

Moreover, results from the OSMlanduse work (Figure 51) showed that the abundance of OSM data, size, distinct features and urban properties were found to be strong factors in defining the categorical accuracy of user observations. Strong confusion was found between agricultural areas (classes 2.1, 2.2 and 2.3). Consequentially, merging those could increase overall accuracy with the side effect of a reduction in thematic depth. Additionally, a connection between small class proportions and low accuracy values could be derived.

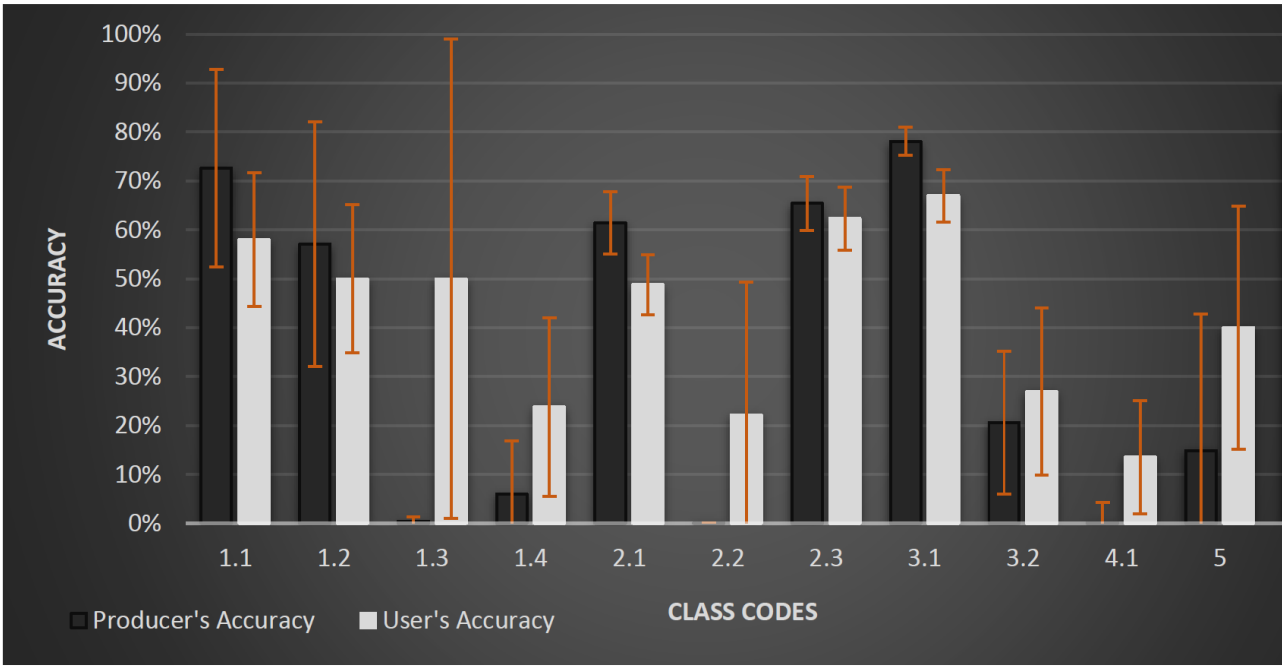

Figure 51: Accuracy estimates of OSMlanduse and a reference data set produced in a mapathon organized with DLR, using the CORINE Land Cover legend: 1.1 = urban fabric, 1.2 = industrial, commercial and transport units, 1.3 = artificial Mine, dump and construction sites, 1.4 = artificial non-agricultural vegetated areas, 2.1 = arable land, 2.2 = permanent crops, 2.3 = pastures, 3.1 = forests, 3.2 = shrub and/or herbaceous vegetation association, 4.1 = inland wetlands and 5 = water bodies

Although this work does not reach the quality levels of regional, tailor-made LULC products, it supports the fast and simple generation of LULC information for any given region, if enough OSM features are present. Some classes are highly suitable for application in LULC while others should be dismissed.

6 Conclusions

This report has built upon the work outlined in deliverable D5.4, which outlined an analysis of the LandSense QA services on selected data collected from the pilot studies. Arising from that deliverable, some modifications, enhancements and new services have been added. A large-scale implementation of the updated LandSense QA platform has been undertaken, with the results presented here. Unlike the preliminary work described in D5.4, the work described here employed all eight of the QA tools identified as necessary and subsequently developed. These tools were applied to all data collected in phases I and II for the applicable pilot studies. Key findings include:

- The QA platform performed as expected, and no notable errors were identified.
- QA checks for polygonal data were carried out successfully. A significant increase in QA failures was identified for phase II data. This will be further investigated in D5.7, and methods will be explored to correct/mitigate issues with polygon overlap.
- The overall performance of the photo privacy checks (90%+ success rate) was very promising. A detailed analysis of the results did highlight some common failure types and potential options for improvement. However, performance concerns with the OpenALPR algorithm used to detect car license plates were identified, which will be further examined in D5.7.
- Photo quality checks for blurring and brightness performed well. A potential correlation between image quality and error rate for photo privacy checks was both hypothesised and detected. Further work, as part of D5.7, will be needed to confirm this hypothesis and explore options for using image quality results to enhance photo privacy checks.
- Results from the pilot studies demonstrated the impact of obstructions to the GPS signal on positional accuracy. How this effect impacts the results is highly dependent on the nature and requirements of a given pilot study and will again feature in the work described in D5.7.
- The contributor agreement results showed high utility for assessing quantitative data, especially where the LULC features to be identified are unambiguous. The results for more qualitative data, such as collected by the MijnPark.NL pilot study, were more limited. Further work will examine options for further improvements and broader measures of contributor agreement.
- The overall results for categorical accuracy were very promising. It was found VGI was of sufficiently good quality for identifying key types of LULC such as residential land use change or detecting and identifying the urban fabric. However, some specific LULC features were harder for VGI to identify accurately, e.g., distinguishing between different types of agricultural land use.

7 References

Bakhtan, M. A. H., Abdullah, M. and Rahman, A. A. (2017) 'A review on License Plate Recognition system algorithms', *ICICTM 2016 - Proceedings of the 1st International Conference on Information and Communication Technology*, (May), pp. 84–89. doi: 10.1109/ICICTM.2016.7890782.

Capellan, S. (2020) D4.7 Demo 3: Forest and habitat monitoring using innovative technologies II. H2020 LandSense

Cohen,W. B., Yang, Z. & Kennedy R. (2010), "Detecting trends in forest disturbance and recovery using yearly Landsat time series: 2. TimeSync - Tools for calibration and validation," *Remote Sensing of Environment*, vol. 114, no. 12, pp. 2911–2924, 2010, doi: 10.1016/j.rse.2010.07.010.

Danylo O, Moorthy I, Sturn T, See L, Laso Bayas J-C, Domian D, Fraisl D, Giovando C, et al. (2018). The Picture Pile Tool for Rapid Image Assessment: A Demonstration using Hurricane Matthew. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences* IV-4: 27-32. DOI:10.5194/isprs-annals-IV-4-27-2018.

Fleiss, J.L.(1971) Measuring nominal scale agreement among many raters. *Psychological Bulletin* 1971, 76, 378–382.

Foody, G. M. (2009) "Sample size determination for image classification accuracy assessment and comparison," *International Journal of Remote Sensing*, vol. 30, no. 20, pp. 5273–5291, Sep. 2009, doi: 10.1080/01431160903130937.

Jain, V. *et al.* (2016) 'Deep automatic licence plate recognition system', *ACM International Conference Proceeding Series*. doi: 10.1145/3009977.3010052.

Jiao, Z. and Fan, H. (2019) 'License plate recognition in unconstrained scenarios based on ALPR system', *ACM International Conference Proceeding Series*, pp. 540–544. doi: 10.1145/3366194.3366290.

Menon, A. and Omman, B. (2018) 'Detection and Recognition of Multiple License Plate from Still Images', *2018 International Conference on Circuits and Systems in Digital Enterprise Technology, ICCSDET 2018*. IEEE, pp. 1–5. doi: 10.1109/ICCSDET.2018.8821138.

Mrkajić, V. (2020) D4.6 Demo 2: Monitoring agricultural land use and provision of value-added agricultural services II. H2020 LandSense.

Nejati, M., Majidi, A. and Jalalat, M. (2016) 'License plate recognition based on edge histogram analysis and classifier ensemble', *2015 Signal Processing and Intelligent Systems Conference, SPIS 2015*. IEEE, pp. 48–52. doi: 10.1109/SPIS.2015.7422310.

Olofsson, P., Foody, G.M., Herold, M., Stehman, S.V., Woodcock, C.E. and Wulder, M.A., 2014. Good practices for estimating area and assessing accuracy of land change. *Remote Sensing of Environment*, 148, pp.42-57.

Liu, L., Jolivet, L., Olteanu-Raimond, A., le Bris, A. (2019) D5.5 Operational workflows for integration of citizen-observed data in authoritative systems. H2020 LandSense. DOI: 10.5281/zenodo.3670937

Rosser, J. Schultz, M. (2019): D5.3 - Quality evaluation of citizen-observed data to the LandSense demonstration cases. H2020 LandSense. DOI:10.5281/zenodo.3670928

Schultz, M. (2018): D5.2 - Definition of citizen-observed and authoritative data collection requirements for LandSense demonstration cases. H2020 LandSense. DOI: 10.5281/zenodo.3670341

Silva, S. M. and Jung, C. R. (2018) 'License plate detection and recognition in unconstrained scenarios', *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11216 LNCS, pp. 593–609. doi: 10.1007/978-3-030-01258-8_36.

Strahler, A. H. et al (2006) "Global Land Cover Validation: recommendations for evaluation and accuracy assessment of global land cover maps," 2006.

Stickler, T. (2020) D4.5 Demo 1: Cost reduction and data conflation in monitoring land change I. H2020 LandSense

Yang, S. *et al.* (2016) 'WIDER FACE: A face detection benchmark', *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016-December, pp. 5525–5533. doi: 10.1109/CVPR.2016.596.

Zhou, Y., Liu, D. and Huang, T. (2018) 'Survey of face detection on low-quality images', *Proceedings - 13th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2018*. IEEE, pp. 769–773. doi: 10.1109/FG.2018.00121.