Validation of the European CORINE Land Cover 2000 database

Gergely Maucha & George Büttner
FÖMI, Budapest, Hungary; European Topic Centre on Terrestrial Environment (ETC-TE), Universitat Autonoma de Barcelona, Spain
e-mail: g.maucha@rsc.fomi.hu

Keywords: CORINE, land cover, LUCAS, land use, validation

ABSTRACT: The CORINE Land Cover 2000 (CLC2000) database, managed by the European Environment Agency (EEA) has been completed in 2004. The database covers 29 countries of Europe. It has been produced by national teams under the supervision of the CLC2000 Technical Team, managed by the European Topic Centre on Terrestrial Environment. Results of the CLC2000 project have been transformed into European projection and various products are available as seamless databases via Internet. CLC2000 (and CLC-change) databases are used to support several policy relevant issues, such as introducing environmental concerns into CAP, the NATURA2000 or the development of environmental indicators. For users of CLC2000 data it is essential to know the accuracy of the database. In case of a European database, a cost-effective final accuracy assessment is not trivial. Validation requires independent, higher resolution data. The subject of the present study is a comparison of CLC2000 database with the LUCAS survey of Eurostat. LUCAS, which is a ground-sampling-based survey, was implemented in 18 European countries in 2001 and 2002. The results of the study have shown that the 85% accuracy requirement specified in Technical Guidelines of CLC2000 had been fulfilled.

1 INTRODUCTION

1.1 *Background*

CORINE Land Cover 2000 (CLC2000) is a project jointly managed by the Joint Research Centre (JRC) and the European Environment Agency (EEA) to update CORINE Land Cover database of Europe for the year of 2000. Landsat-7 ETM satellite images acquired within the frames of the IMAGE2000 project were used for the update. The first CORINE Land Cover (CLC) inventory in EU15 and most of the new member states was implemented between 1985 and 1996 with the aim to characterise land surface by using a uniform nomenclature across Europe at a scale of 1:100.000. The CLC nomenclature includes basically land cover items, but land use elements are also found in it, especially for built-up environment. The CLC database is a digital map, covering countries with a seamless polygon database with the minimum mapping unit (MMU) of 25 ha. The CLC nomenclature includes 44 categories in five major groups (Heymann et al. 1994; Bossard et al. 2000).

The European Land Use/Cover Area Frame Statistical Survey (LUCAS) is a project managed by the Statistical Office of the European Communities (Eurostat). It has the main purpose of providing harmonised information on agri-environment of Europe. LUCAS, as implemented in 2001-2003, recorded land use and land cover information in a two-level regular grid. The grid size was 18×18 km (primary sampling units, PSU). Each PSU included 10 secondary sampling units (SSU). SSUs

were placed in two parallel rows, with five points in each row. The distance between SSU points was 300 meter. The field surveyor located the point with high accuracy (specified as 1–3 meter) and registered LC and LU information – and other environmental variables – related to this point, according to strict guidelines (Avikainen et al. 2003). The area of observation was usually a circle of 1.5-meter radius, in exceptional cases a circle with 20 m radius had to be considered. Digital field photographs were taken in East, West, North and South directions usually in each central SSU point of the first row. The LUCAS LC nomenclature includes 57 categories in 7 major groups, while LUCAS LU codification applies 14 land use classes (Duhamel et al. 2003).

1.2 *Objectives*

The objective of validating geographical data is to derive final accuracy/reliability figures using independent, higher resolution and more accurate data that are related to a similar data acquisition period and not used in the compilation of the target database. Some countries have performed a national validation of their CLC2000 database, using country specific approaches. In most cases spatially and thematically more detailed data were used for that purpose (G.W. Hazeu 2003; Caetano & Mata. 2005). Validation in this context means thematic validation, as the geometry of CLC2000 is much more trivial to check by comparing with IMAGE2000. The general aim of validating CLC2000 is to assess its thematic accuracy by means of a statistical method applied homogeneously for as many as possible participating countries.

LUCAS data are the only information that is available in many countries for European-wide validation of CLC2000. LUCAS fulfils the criteria of validation data due its following characteristics:

- higher geometric accuracy,
- higher thematic accuracy (LC/LU data based on field observations, or in exceptional cases on photo-interpretation of aerial ortho-photographs),
- data acquisition is nearly coincident,
- independent, as usually not used during production of CLC2000.

The specific questions of the validation were:

- Is the 85% target accuracy of CLC2000 data (EEA-ETC/TE, 2002) fulfilled?
- Which CLC classes are determined with higher accuracy and which ones with lower accuracy?
- What are the reasons for misinterpretations?

Due to lack of similar reference data with wide coverage for the period of the first CLC inventory (1985-1996) the other two deliverables of the CLC2000 project (CLC-change and revised CLC90) were not validated.

2 DATABASES

Three European databases were used in the validation process.

2.1 *CLC2000 data*

29 European countries have finished CLC2000 (EU25, Bulgaria, Croatia, Liechtenstein and Romania). National Teams produced CLC2000 databases, based on IMAGE2000 satellite imagery. Most countries used computer-assisted photo-interpretation. Finland, Sweden and UK applied a more computer-oriented approach, tailored to the geography and to the databases available in these countries.

On behalf of the EEA, the European Topic Centre on Terrestrial Environment (ETC-TE) was responsible for quality control, quality assurance and management of national CLC2000 databases and for production of European deliverables. Geometric precision (localisation accuracy) of CLC2000 is better than 100 meter, according to specification. Planned overall thematic accuracy is better than 85% (EEA-ETC/TE, 2002; Büttner et al 2004.). CLC2000 data are available for the public on the EEA data server (http://dataservice.eea.eu.int).

2.2 *IMAGE2000 data*

IMAGE2000 data means ortho-corrected Landsat-7 ETM satellite imagery, taken in the period of 1999-2001 (2002) covering countries participating in the CLC2000 project. IMAGE2000 data were produced on a centralized way coordinated by JRC. IMAGE2000 data used for validation were received in national projection from EEA. Geometric accuracy of the IMAGE2000 data has been specified as better than 25 m RMSE. IMAGE2000 data are available for the public on JRC data server (http://image2000.jrc.it/i2000).

2.3 *LUCAS 2001–2002 data*

The first national LUCAS surveys were implemented in 13 EU member states in 2001. UK and Ireland were left out because of the foot and mouth disease. These two countries were surveyed in 2002 together with three (that time) accession countries (Estonia, Hungary, and Slovenia). Eurostat was responsible for coordination of national LUCAS surveys. Comparison of some elements of CLC and LUCAS databases can be found in Table 1. LUCAS data covering EU15 have been received from Eurostat for the purposes of this validation. Data covering the three other countries have been provided by the responsible national Statistical Offices.

Table 1. Comparison of main parameters of CLC2000 and LUCAS 2001-2002.

	CORINE Land Cover	LUCAS
Coverage	EU25, Bulgaria, Croatia, Liechtenstein, Romania	EU13 (2001); United Kingdom, Ireland, Estonia, Hungary and Slovenia (2002)
Database characteristics	Single land cover database with land use elements	Separate land use and land cover databases; field photographs
Method of database production	Mapping, based on satellite images and topographic maps	Area frame sampling in 18 × 18 km grid (PSU), 10 sampling point in each grid elements (SSU)
Nomenclature	Land cover (with some land use elements); 44 classes in 5 major groups	Land cover: 57 classes in 7 groups, emphasis on agriculture
Observation unit	25 ha (250 000 m^2) MMU, 100 m minimum width	Land use: 14 classes in 4 groups 7 m^2 (0.1 ha = 1000 m^2 in exceptional cases)
Coding	Single level-3 code in each polygon	One (or two) LC code, one (or two) LU code in each SSU + additional environmental information
Localisation accuracy	Better than 100 meters	1-3 meter

The area of 18 countries participating in LUCAS together represents 77.5% of the total area of 29 countries participating in CLC2000. Larger lakes, high mountains, far-Northern areas were not sampled by LUCAS.

3 METHODOLOGY

For validation of CLC2000, two different methods were applied, which measure different kinds of accuracy.

The methods used were as follows:

- Automatic comparison of CLC2000 and LUCAS LC/LU codes.
- Reinterpretation of IMAGE2000 with the help of LUCAS data (LU & LC codes and photographs) following interpretation rules of CLC.

3.1 *Setting up the working environment*

In order to be able to view and handle LUCAS data together with CLC data, a GIS environment was set up under ESRI ArcView 3.2 software. Main steps of this process were as follows:

- LUCAS data tables provided in MS/Access database (EU15) or a formatted text file (EE, HU, SI) (M. Kayadjanian 2001) were imported into ESRI database format and were joined to coordinates using ID of SSU points;
- The position of LUCAS SSUs was transformed from ETSR89 into the national projection in all 18 countries;
- Landscape photographs were joined to the coordinates of the corresponding SSU point, so photos could be displayed automatically based on SSU IDs;
- The CLC2000 database and the IMAGE2000 database were prepared for easy, simultaneous display with LUCAS data. In order to reduce data volume, 20 km × 20 km imagettes around LUCAS PSUs were subset from IMAGE2000.

3.2 *Automatic comparison*

The nomenclatures of CLC and LUCAS although similar, are not directly compatible. To resolve this, a correspondence table between CLC and LLC classes was created. In this table only basic constituents of each CLC class were considered and not those additional elements, that might be present in a given polygon because of the generalisation. As the CLC is not a pure land cover nomenclature – especially considering the 1st group of classes (artificial surfaces) – the LLU information also had to be considered during comparison. Therefore a similar table was constructed between CLC and LLU.

Table 2. Examples of correspondence between CLC and LUCAS classes.

CLC code	CLC class name	Corresponding LLC codes	Corresponding LLU codes
112	Discontinuous urban fabric	A11 (Buildings with 1 to 3 floors), A12 (Buildings with more than 3 floors), A2 (Non built up area and linear features), B43 (Other fresh vegetables), B44 (Floriculture and ornamental plants), C (Forest and tree areas), E (permanent grassland)	U11 (Agriculture), U31 (Transport, communication, storage, protective works), U34 (Commerce, finance, business), U35 (Community services), U36 (Recreation leisure, sport), U37 (Residential)

Using the correspondence table the automatic comparison included two steps:

- GIS overlay of all secondary sampling units (SSU) and CLC2000.
- Comparison of codes using the correspondence table and calculation of the degree of agreement between the two databases. Each SSU was checked, whether the CLC code corresponded to (one of) the LLC class(es) in the correspondence table. Land use was handled in a similar way. The value of agreement was considered positive, if both the LC and LU values corresponded to the CLC data according to the table. Finally, the agreement cases were summed and the percent total agreement (PTA) was computed.

The PTA figure shows to what extent CLC2000 reflects the reality, which is modelled by LUCAS LC and LU codes. Benefits of this method are fast, automatic computation and the inclusion of a large number of points (around 100 000 SSUs).

3.3 *Evaluation based on reinterpretation of LUCAS photographs*

At least one LUCAS landscape photograph is available for 8231 locations out of the 10258 PSUs observed in 18 countries implementing both LUCAS and CLC2000. Validation is based on a visual

re-interpretation of IMAGE2000 around the LUCAS PSU, using also the reference information provided by LUCAS landscape photographs, moreover the LLC and LLU information for the corresponding SSUs.

The main advantages of this method are:

– CLC generalisation rules and complex class definitions can be considered;
– LLC and LLU codes in the 1200 × 300 meter area surrounding the central SSU (where the photos were taken) provide the spatial context;
– LUCAS field photographs provide a valuable "ground truth" for re-interpretation.

Results provided by this method can be considered as thematic reliability figures of CLC2000, as national CLC2000 data are compared to an independent CLC2000 interpretation.

The re-interpretation included the following three steps:

1. The validation expert provided the CLC code by visual interpretation of the location around the LUCAS sampling point, where the LUCAS landscape photographs had been taken. The interpretation was based on IMAGE2000 data, LLU and LLC data and landscape photographs. Additional environmental information collected by LUCAS could also be consulted (e.g. irrigation, natural hazards). No information from the CLC database was shown at this moment. The validating expert had to respect all the CLC interpretation rules (MMU, class definitions, generalisation, etc) when providing the code. After confirmation the code could not be changed.
2. The CLC polygon boundaries were also displayed on top of satellite image and LUCAS points, but still without class information. This situation was interpreted again by assigning a second CLC code. After confirmation the code could not be changed.
3. The actual CLC code (taken from the CLC2000 database) was also displayed. By comparing the actual CLC code and the control codes the validating expert had to evaluate the situation as one of the four following cases:
   'Clear agreement' was used when the agreement between the control interpretations and the CLC2000 was without doubt.
   'Another interpretation is possible' was used when the interpretation was still acceptable, but the control codes did not fully agree with the CLC2000 code. The validating expert had to select the explanation whether: another delineation was possible; another code was possible; or sampling point fell close to LC boundary.
   'Wrong interpretation' was used when the CLC2000 interpretation was not acceptable. The validating expert had to select the explanation whether: wrong code; wrong generalisation; not enough details (omission); or inaccurate delineation. Inaccurate delineation means if matching of IMAGE2000 and CLC2000 deviates with more than 100 meters.
   'Not enough information' was used in those rare cases, when the available information was insufficient to make a well-based decision (e.g. cloudy IMAGE2000, bad quality LUCAS photos, etc.).

3.4 *Estimation of the representativity*

When looking for the reliability estimates of CLC2000 using LUCAS data, we have to consider the representativity of the estimates. The representativity was calculated based on the assumption of a binomial distribution, on the confidence level of 95% (±2 standard deviation). The representativity of the estimation depends on the size of the class and the number of LUCAS samples available for validation. In case of a small CLC class the small number of LUCAS samples might provide misleading results, as the error of the estimation is high.

The binomial distribution is a good approximation for the PSU level LUCAS sampling (reinterpretation), but it is distorted on the SSU level (automatic comparison), because of the two-level sampling method – the samples cannot be considered totally independent. A case study performed on Hungarian data showed (Maucha et al. 2003) that on SSU level the standard deviation of the resulting distribution could be estimated by doubling the standard deviation of the binomial distribution (corresponding to the total number of samples).

4 RESULTS

4.1 *Results of automatic comparison*

A correspondence table between CLC and LUCAS LC&LU was constructed in the frames of the pilot study (Maucha et al. 2003) and was refined using the results of the cross-tabulation. CLC2000 and LUCAS LU&LC covering 18 countries were compared and analysed using 99.936 SSUs by means of the correspondence table. As no LUCAS LC class exists for sea water, moreover LUCAS usually does not sample the sea, CLC class 523 (sea and ocean) and 423 (intertidal flat) were omitted from this analysis. Summary results of the automatic comparison are shown in Table 3.

Table 3. Summary results of the automatic comparison of CLC2000 and LUCAS.

No of SSUs	LC OK	AE	LU OK	AE	PTA (LC&LU OK)	AE
99 936	80.1%	±0.5%	88.7%	±0.4%	75.6%	±0.5%

Main results of the automatic comparison are as follows:

- Percent total agreement (PTA) between CLC2000 and LUCAS LU&LC is 75.6 ± 0.5%. This value can be interpreted as follows: CLC2000 approximates the reality with 75.6% average accuracy (i.e. out of four randomly selected locations three samples are expected to be correctly classified by CLC2000). Considering the large minimum mapping unit of CLC2000 (25 hectares) these results can be considered to be rather satisfying.
- Looking at the class-level values, 13 of the 44 CLC classes could not be tested because of low representativity. By removing the 2.4. × classes (complex agriculture) as well from the analysis (as having no meaning at the SSU level), altogether 27 classes were validated.

4.2 *Results of reinterpretation*

Using the methodology introduced under 3.3, altogether 8231 PSUs were analysed in 18 countries. "Clear agreement" and "Another interpretation is possible" cases together yield the number of samples with acceptable CLC2000 classification (7058 cases). This makes up 87% of all samples used in the validation (8115 cases). Summary results of the reinterpretation are shown in Table 4.

Table 4. Summary results of the reinterpretation of IMAGE2000 using LUCAS data.

No of PSUs	Clear agreement	Wrong interpretation	Another interpretation possible	Not enough information	Samples used for validation	Reliability of CLC2000	AE
8 231	6 681	1 057	377	116	8 115	7 058	±61
100%	81.2%	12.8%	4.6%	1.4%	100%	87.0%	±0.8%

Main results of the reinterpretation approach are as follows:

- The total reliability of CLC2000 is 87.0 ± 0.8%. This value is based on an independent CLC interpretation performed around LUCAS PSUs. Therefore it can be concluded, that the 85% accuracy requirement specified in Technical Guidelines (EEA-ETC/TE 2002) has been fulfilled.
- Looking at the class-level values, 22 of the 44 CLC classes could be validated (relative error less than 50%), i.e. 22 CLC classes could not be validated because of low representativity (or intentional omissions by LUCAS, e.g. glaciers or sea). Classes impossible to validate belong mostly to artificial surfaces, wetlands and water.
- The two largest CLC classes in the 18 countries investigated are arable land (211) and coniferous forest (312). Together with two other agricultural classes: agro-forestry (244) and permanently irrigated land (212), these were estimated with high reliability (90-95%).

– The lowest class-level reliability (below 70%) have been obtained for sparse vegetation (333), highlighting the difficulties in interpreting this class.
– As shown by Figure 1, the highest class-level reliability (>95%) was obtained for rivers (511), lakes (512), industrial and commercial units (121) and discontinuous urban fabric (112).
– Majority (78%) of classification errors occurred on level 3 and level 2.
– Level-1 misclassifications mostly occured between 'agriculture' and 'forest and semi-natural' classes.

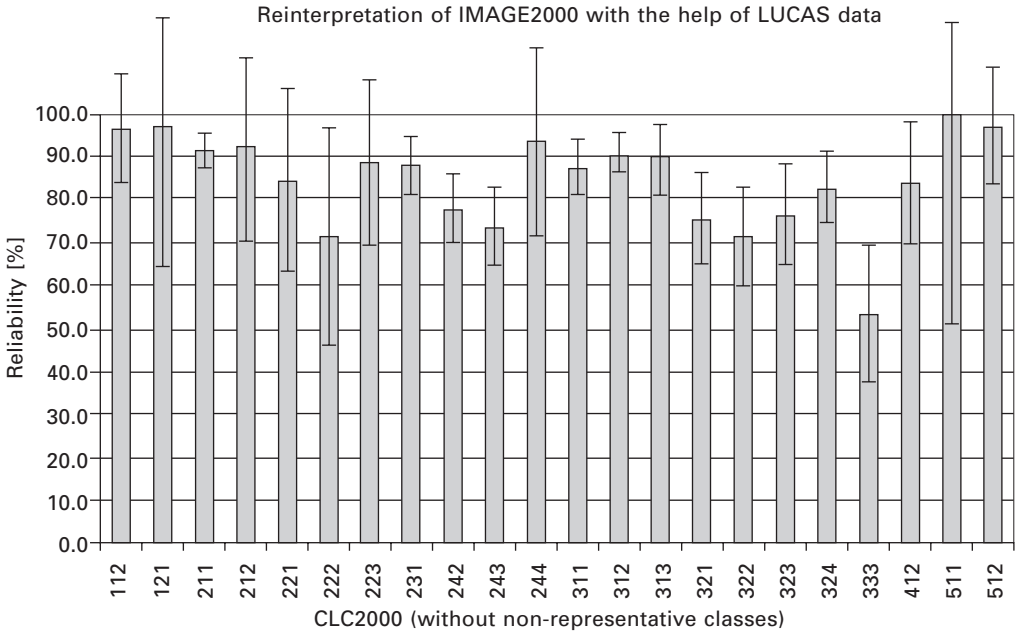

Figure 1. Re-interpretation of IMAGE2000 with the help of LUCAS data: class-level accuracies.

4.2.1 *Analysis of misinterpretations*

Cases classified as "wrong interpretation" were further analysed. Using a set of standard explanations (see Ch. 3.3), the 1057 cases were labelled as seen on Table 5.

Table 5. Summary results of the analysis of misinterpretations.

All cases of "wrong interpretation"	Wrong code	Not detailed interpretation	Wrong generalisation	Inaccurate delineation
1057	472	435	134	15
100%	44.7%	41.2%	12.8%	1.8%

"Wrong code" (= commission errors) and "not detailed interpretation" (= omission errors), being the two main reasons of mistakes, showed a similar number of occurrences. Drawing accuracy (i.e matching of IMAGE2000 and CLC2000) was the less important source of error. This reflects the many efforts to improve the geometric accuracy of CLC90.

5 CONCLUSIONS

CLC2000 data covering 18 countries of Europe (3.4 M km^2) have been validated by means of LUCAS data. Two methods were applied:

– Automatic comparison of CLC2000 codes and LUCAS LU & LC codes over more than 100.000 SSUs;
– Reinterpretation of IMAGE2000 data around more than 8200 LUCAS PSUs based on ground photographs and LUCAS LU & LC codes.

Accuracy/reliability figures for the total amount of samples and also on class level have been derived. Representativity (error values) of the estimates has also been derived based on statistical principles.

The main results of the validation are as follows:

– The use of LUCAS data was an appropriate tool in the validation process of CLC2000 on a European level (18 countries).
– The reinterpretation method proved to be an appropriate way of measuring the reliability of CLC interpretation. The total reliability of CLC2000 stated by this method is 87.0 ± 0.8%, concluding that the 85% accuracy requirement specified in Technical Guidelines (EEA-ETC/TE, 2002) has been fulfilled.
– The two main sources of mistake were misclassification and not enough detailed interpretation. Delineation accuracy was a less important source of error.
– At class-level, reliability of half of the CLC classes could not be validated because of low representativity. Classes not possible to validate belong mostly to artificial surfaces, wetlands and water.
– The automatic comparison proved to be an appropriate method to estimate the correspondence between CLC interpretation and the reality. According to this study, the CLC2000 database approximates the reality with 75.6% ± 0.5% average accuracy

ACKNOWLEDGEMENTS

The CLC2000 project is part of the work programme of the European Topic Centre on Terrestrial Environment (ETC-TE) working under contract with the European Environment Agency between 2001-2005. The authors express their gratitude to Chris Steenmans EEA project manager, Adriana Gheorghe EEA-EIONET coordinator and Stefan Kleeschulte ETC-TE manager for continuous cooperation and support. Thanks also to Vanda Lima and Javier Gallego (JRC) for useful comments on the validation methodology. The leading photo-interpreter of the Hungarian CLC2000 team, Mária Bíró performed the reinterpretations. Special thanks to László Oszkó for his help on statistical considerations. Linguistic corrections were provided by Barbara Kosztra.

REFERENCES

Avikainen, J., Delincé, J., Croi, W., Kayadjanian, W., Bettio, M. & Mariano, A., 2003. LUCAS Land Use/Cover Area Frame Statistical Survey. Technical Document No. 1: Sampling plan. (version 2.4) EUROSTAT/LAND/LUCAS1.

Bossard, M., Feranec, J. & Otahel, J., 2000. CORINE Land Cover Technical Guide – Addendum 2000. Technical report No 40. Copenhagen (EEA). http://terrestrial.eionet.eu.int.

Büttner, G., Feranec, J., Jaffrain, G., Mari, L., Maucha, G. & Soukup, T., 2004. The CORINE Land Cover 2000 Project. EARSeL eProceedings 3(3), 331-346.

Caetano, M. & Mata, F., 2005. Accuracy assessment of CLC2000 for Portugal. 25th EARSeL Symposium, Porto, 6-9 June 2005; (to be published in the proceedings).

Duhamel, C., Eiden, G., Aifantopoulou, D. & Croi, W., 2003. LUCAS Land Use/Cover Area Frame Statistical Survey. Technical Document No. 2: The Nomenclature. (version 1.5) EUROSTAT/LAND/LUCAS2.

EEA-ETC/TE, 2002. CORINE Land Cover update, I&CLC2000 project, Technical Guidelines, http://terrestrial.eionet.eu.int.

Hazeau, G.W., 2003. CLC2000. Land Cover database of the Netherlands, Alterra Rapport, Wageningen, 775 ISSN 1566-7197.

Heymann, Y., Steenmans, Ch., Croissille, G. & Bossard, M., 1994. CORINE Land Cover. Technical Guide. EUR12585 Luxembourg Office for Official Publications of the European Communities.
Kayadjanian, M., 2001. LUCAS Land Use/Cover Area Frame Statistical Survey. Technical Document No. 6: Data transfer and control procedures. (version 2.2) EUROSTAT/LAND/LUCAS6.
Maucha, G., Büttner, G. & Kosztra, B., 2003. Applying LUCAS data for verification/validation of CLC2000. A pilot study for Hungary. (ETC-TE internal report).