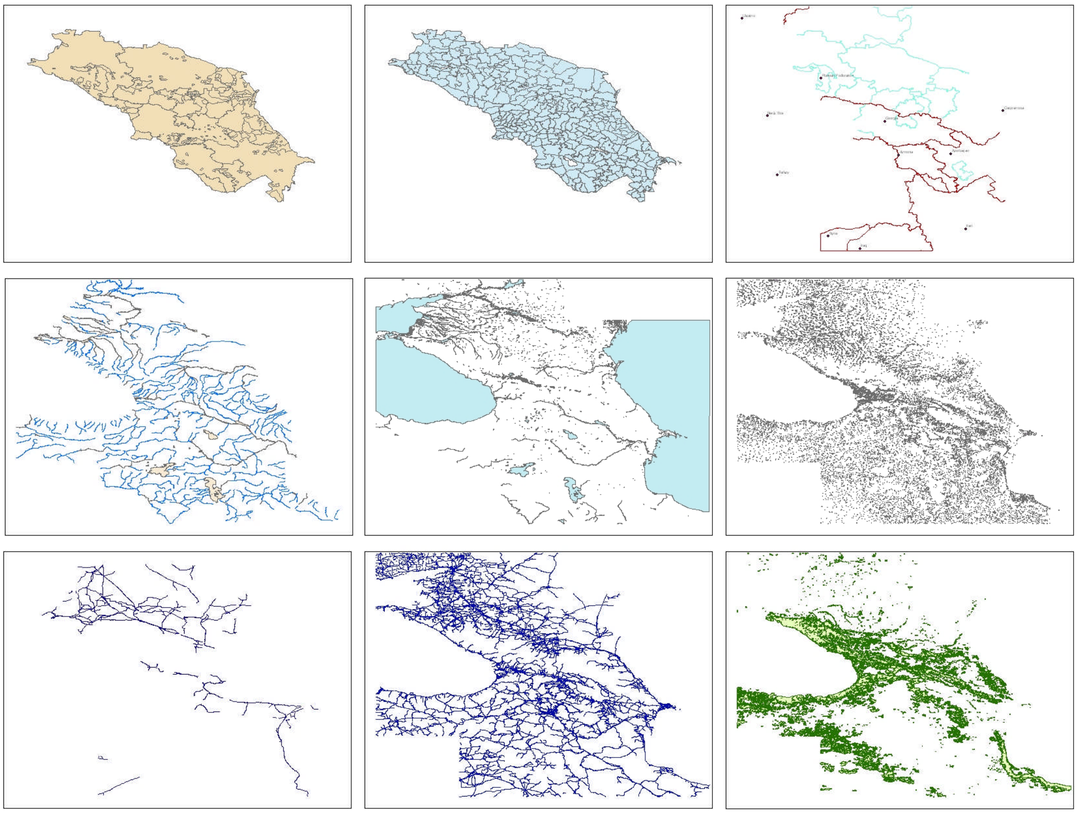
JRC Scientific and Technical Reports

Building a spatial decision support system for conflict prevention in the Caucasus

Harmonization of heterogeneous sources and data quality assessment procedures

Stephenne Nathalie
MacDonald Chris



EUR 22949 EN - 2007

JRC
EUROPEAN COMMISSION

Institute for the Protection
and Security of the Citizen

The Institute for the Protection and Security of the Citizen provides research based, systems-oriented support to EU policies so as to protect the citizen against economic and technological risk. The Institute maintains and develops its expertise and networks in information, communication, space and engineering technologies in support of its mission. The strong cross fertilisation between its nuclear and non-nuclear activities strengthens the expertise it can bring to the benefit of customers in both domains.

European Commission
Joint Research Centre
Institute for the Protection and Security of the Citizen

Contact information
Address: EC -JRC – IPSC - "Support to External Security"
TP 267, Via Fermi 1
21020 Ispra (VA), Italy
E-mail: nathalie.stephenne@jrc.it
Tel.: +39 033278 6539
Fax: +39 033278 5154

http://ipsc.jrc.ec.europa.eu
http://www.jrc.ec.europa.eu

Legal Notice
Neither the European Commission nor any person acting on behalf of the Commission is responsible for the use which might be made of this publication.

JRC 40597

EUR 22949 EN
ISBN 978-92-79-07140-9
ISSN 1018-5593

DOI: 10.2788/39806

Luxembourg: Office for Official Publications of the European Communities

Printed in Luxembourg

1. Table of Content

2. Abstract

Geodata analysis at regional level integrates inevitably some datasets from various sources (statistical, geographical, environmental,…), various scale (regional, national, ..) and various quality: While political structures are constantly changing, as in a potentially conflicting region such as Caucasus, these data integration issues increase. Implementation of quality control methods is an initial and essential step in the integration of geodata inside a spatial regional model. This report provides tools for data harmonization that can be applied to other datasets and other region when existing data sources do not evaluate the quality of their information.

The goal of this report is to provide a quality assessment of the Caucasian GIS dataset to build the Caucasus geomodel of instability/stability. This report evaluates qualitatively and quantitatively the adequacy of this dataset to the objective in following a structured quality assessment protocol (Johnston et al. 1999) and consolidates a final geodatabase. Integrating data from a multitude of derivative geospatial products produced by different sources pose severe problems. Challenges are also introduced by the GIS technology itself. Various data are introduced in this study but the main source of statistical and spatial information is the acquisition of the geopolitical atlas dataset, the "Caucasian dataset" (Radvanyi, INALCO, 2006).

In this report, four data quality elements are identified and described in the specific case of the Caucasian dataset. Lineage information, the three accuracy dimensions (positional, temporal and attribute), logical consistency and completeness evaluations are qualitatively and quantitatively assessed by various metrics. This paper illustrates the use of automatic cartographic and data cleanup techniques of Geographic Information System (GIS) to solve data issues (self overlapping, dangles, pseudonodes and gap in spatial data). This report can further be used as a reference for both the producer and the user to somewhat replace the missing metadata information. Clear statements on dataset quality allow to better communicate in a common goal of understanding the geopolitical Caucasus context.

The bulk of this report has aimed to illustrate how spatial data from various sources have been collected and made ready for use within a GIS. The different evaluation tests allow to give an overall estimation of the dataset quality. This type of data cannot be used at a scale higher than approximately 1:500 000. This Caucasian dataset has the objective to provide an overall picture of the regional security complex and not a precise localisation of specific real features. This fact has to be kept in mind in the following processing modelling stages.

Based on the results of this report, especially the completeness and fitness of the dataset to represent the scope of the model, the Caucasus study will further explore two distinct modelling approaches: (i) a spatial and continuous muticriteria model of instability integrating in a continuous GIS the geopolitical factors, (ii) defining instability indicators values for subnational spatial entities (district units) throughout the Caucasus region.

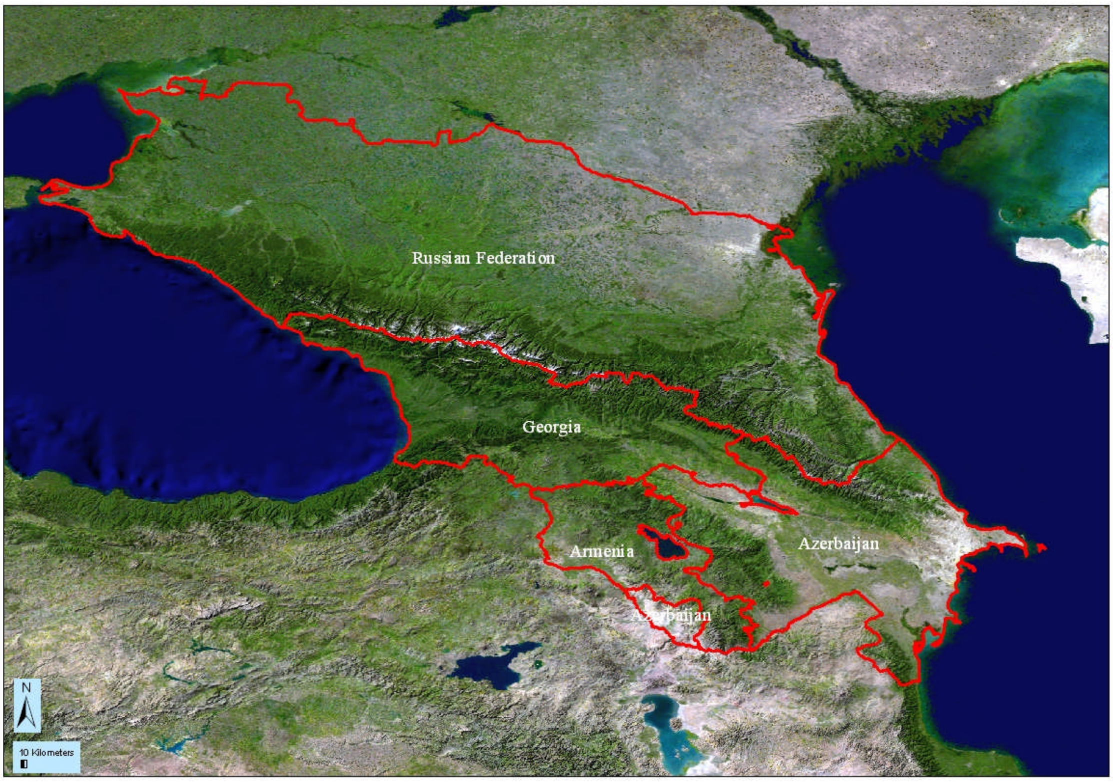
This report provides an adapted methodology to assess quantitatively the quality of a database when no metadata information is available. The elements of data quality are envisaged in a progressive way in this report and thoroughly studied for the settlement layer. The other layers are evaluated in a less in-depth way but allow the test of different methods associated to the three types of features (point, line, polygon).

3. Introduction

Combining geographic information systems and modelling techniques create opportunities to better understand, analyse and support the management of instability and security. Only a spatial quantitative simulation model allows a holistic representation of the instability driving forces to generate “what-if” scenarios exploring the influence of single or groups of variables on the regional security. Based on clear assumptions, an instability model can be useful for investment prioritization, and simulation of the impacts of different political decisions. Simulation models emphasize the interactions among the components of the system and take into account the quantitative effects of each driving force.

JRC [1] started to work on this integrated approach within the context of Neighbouring Countries, on a particular complex security region : the Caucasus (Figure 1 : Caucasian study area (MODIS Rapid Response System, Copyright NASA). The geomodeling on the Caucasus aims to provide a holistic view that will include environmental issues related to the security of the region through a quantitative analysis by means of a Spatial Decision Support System (SDSS). The first step of this approach consisted in collecting information and data relevant to conflict in the Caucasus region.

Figure 1 : Caucasian study area (MODIS Rapid Response System, Copyright NASA)



Contacts and cooperation with relevant organizations and/or research institutions working on and in the region (UNEP, OSCE, Institut National des langues et civilizations orientales, Paris) were established. With the objective of building a geopolitical atlas of the Caucasus, a comprehensive GIS dataset on the region, called “Caucasian dataset” in the text below, has been developed. The producers are J. Radvanyi, director of the Observatoire des Etats post-soviétiques, at the Institut National des Langues et des Civilisations (INALCO, Paris) in a close collaboration with Nicolas Beruchashvili,

[1] Information support for Effective and Rapid External Action (ISFEREA) Action n° 41002, in support of DG RELEX and its EU Neighbourhood Policy, within the framework of an Administrative Arrangement DG RELEX/ DG JRC N0 05-0042 on scientific and geoinformation system and IT based support to EC Crisis Response of Rapid Reaction Mechanism (RRM) : Work Package 2006 - Item 6

Georgian professor in Geography (Cartographic Department at the State University of Tbilisi) and other professional cartographers thorough the Caucasus region as Vitaly Belozerov, Ashot Khoetsian and Musseib Musseibov who are respectivly Head of the Departement of Geography at Stavropol (Russia, North Caucasus), Erevan (Armenia) and Bakou (Azerbaijan) State University. In 1997, a first hand-made Atlas of the geopolitical context of Caucasus has been edited by Radvanyi and Beruchashvili using this dataset in Mapinfo environment. These authors are currently working on an update of this Atlas, including a JRC collaboration. While the local contacts and the integration of geographical and security focus give an extraordinary level of interest to this dataset, data have been provided without metadata and quality level.

Geodata analysis at regional level involves the integration of dataset from various sources (statistical, geographical, environmental,…) and various quality, especially in a potentially conflicting region such as the Caucasus where the political structures are constantly changing. .Implementation of quality control methods is an initial steps in the integration of geodata in a spatial regional model.

The goal of this report is to provide a quality assessment of the geospatial data assembled for developing a geomodel of instability. This report refers to a quality assurance protocol designed for the US Army (Johnston et al. 1999). Following and adapting the structure of the protocol, this report describes the original “Caucasian dataset”, as it has been provided by the regional experts, but also the technological issues of using these data in a modelling study. This report proposes solutions to overcome some issues and to consolidate an improved version of a geospatial database.

The resolution and data quality needed for geomodelling differ from a geopolitical mapping exercise. This paper is divided in three main sections : explanation of the adapted methodology used to assess the quality of data available (Section 4), a description of the original dataset and its quality issues (Section 5), an explanation of consolidation steps carried out either in the topological data hierarchy or in the statistical coherence (Section 6). This paper illustrates the use of automatic cartographic and data cleansing techniques of Geographic Information System (GIS) to solve data issues.

4. Quality assessment method

Integrating data from a multitude of derivative geospatial products produced by different sources pose severe problems. Challenges are also introduced by the GIS technology itself. Error propagation are related to the interoperability between data representation, computer hardware or software problems and data processing (Burrough 1986). Most of these reasons explain most of the dataset issues that will be reviewed herein. Quality assessment and quality control combine theoretical work on the nature of geospatial data and specific assessment of the dataset. The procedure applied in this report is adapted from the one described and tested to a specific dataset (Fort Hood, TX ITAM GIS) by Johnston et al. (1999) for the US Army (web access : http://www.gis.uiuc.edu/research/spatialanalysis/quality%20assurance.htm, on 15 May 2007). Theoretical precisions are found in Longley et al (1999).

Description and analysis of geospatial data quality refer to standards and characteristics documentation defined by organizations and research communities to promote interoperability of these data. Metadata standards (i.e. ISO or FGDC) describe the data file format and accuracy. Unfortunately, these metadata standards are still specific to the different organizations. The US Federal Geographic Data Committee (FGDC) has defined 7 components of geospatial data:

1. Identification – name, developer, geographic extent, thematic types, currentness.
2. Data quality – accuracy elements.
3. Spatial data organization – spatial model, number of objects, encoding methods.
4. Spatial reference coordinate systems, datums, conversion parameters.
5. Entity and attribute information – definitions, content description, coding/representation standards.
6. Distribution – format, media, price, location for obtaining data.

7. Metadata reference – developer, date compiled.

Data quality is not a well-defined concept in the geospatial research. Data quality cannot be described with a single element or figure. In this report, four elements are distinguished (adapted from Longley et al. 1999):

1. The lineage stats on the history of the dataset. This qualitative documentation includes identification of the producer, assumptions, source of observation, compilation methods, transformations or derivations in the process of developing the data set.
2. The accuracy can be divided in three dimensions :
   - Spatial or positional accuracy describes the degree of discrepancy between position of the objects in the dataset and the objects' actual position (measured in the field) or an accepted representation of its objects (another dataset of recognized higher accuracy). While metrics are well defined for point entities but less for lines and polygons. The positional accuracy refer to an horizontal precision in x and y dimensions and should provide a quantitative statistic representing the likely nearness to true position, as RMSE (root mean square error). RMSE is described by empirical frequencies, means and standards deviation of positional errors (Veregin 1999). In vector-based GIS, the epsilon band is defined by a minimum buffer width around the reference object.
   - Temporal accuracy is often associated to the currentness (up to date or not) but in fact calculates the agreement between encoded and "actual" temporal coordinates. It refers to the lineage but is applied to all objects of the database (date of the construction of the spatial object) or to their attribute (date of the survey).
   - Thematic/attribute accuracy is usually assessed in a sample of point location with an error matrix, in analogy of remote sensing classification assessment metrics as overall accuracy and Kappa.
3. Completeness refers to the relationship between the objects represented in a database and the universe of all objects. The definition of the completeness is linked to the role of the information with respect to fitness of use but also to the semantic accuracy. This measure is application dependent and then differ between the provider and the user. A quantitative assessment needs a selection of criteria, definition and mapping rules, the calculation of deviations from standard definition and discrepancy measure between objects.
4. Consistency describes the structural integrity of a data set and the interrelationships between data and attributes. Graphic rules for spatial reference method are for example prohibitions on intersections, nodes, minimum/maximum length area. This assessment does not require a control dataset but real world constraints and approved procedure.

5. Caucasus dataset assessment

These four elements of data quality are used thereafter to analyse the Caucasian dataset in a structured way. This section describes the data and try to quantify their quality by a comparison with other dataset. The following section (See chapter 6) will explain the improvement of the dataset and its consolidation based on the correction of the errors identified in this section. Resulting from a regional mapping exercise, the original Caucasian data has been provided without metadata. Only the existing knowledge of the database manager and data file inspection (inferring metadata from observable characteristics) can be used in this study to assess the quality of the dataset. The results of this process largely depends on the involvement of data producers in the production of the technical report set up by the JRC.

5.1. Lineage

Because of missing direct documentation on the dataset characteristic, a detailed inspection of the original spatial data is conducted to determine lineage. Determination of dataset characteristics relies heavily on inferences based on comparison between the dataset and possible parent material.

5.1.1. Producer

A large regional team has gathered various source of information to build this dataset. This team integrated researchers belonging to four countries (Russian Federation, Azerbaijan, Armenia and Georgia) with extremely different geographical and statistical background. Nicolas Beruchashvili, physical geographer, was the coordinator of the topographical dataset while Jean Radvanyi, social geographer, coordinated the statistical gathering process.

5.1.2. Geographical extent

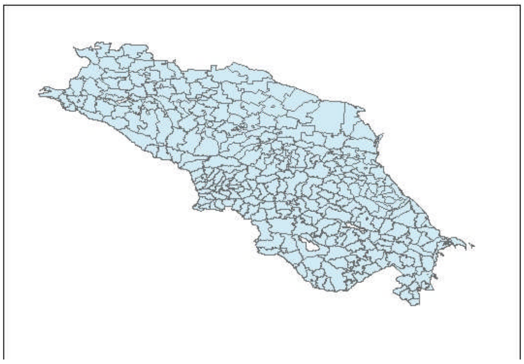
The spatial extension vary within the different themes. In a minimum common coverage, the dataset are covering three countries (Azerbaijan, Armenia and Georgia) and the southern part of the Caucasian region in Russian Federation. While topographical elements extent to neighbouring countries (Turkey and Iran), the attributes of the statistical dataset cover exclusively the three independent countries of the Caucasus and the neighbouring regions in Russian Federation (Figure 1). The units of this socio-economic dataset is NUTS3 in the independent countries and NUTS4 in the Russian part of the region . This thematic choice provides a homogeneity in the size of the statistical units that is interesting for the producer as well as for the user goals. The spatial extent of this statistical dataset is an indication of the main data’s source that is the Census of the national statistical Committees.

5.1.3. Entity types

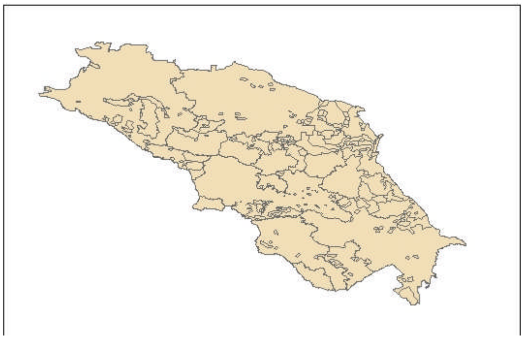
Table 1 shows the eleven entity types of the Caucasian dataset with a quick look to illustrate their variable extents.

Table 1 : Topographical themes

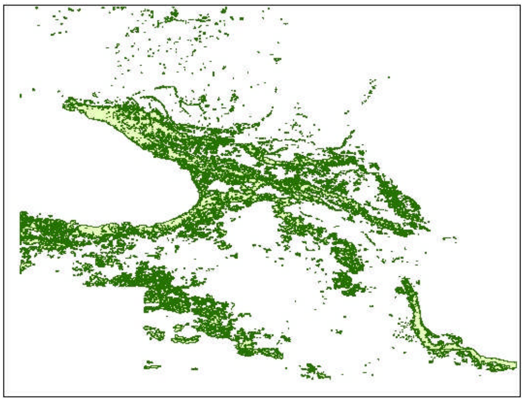
District borders (polygons)



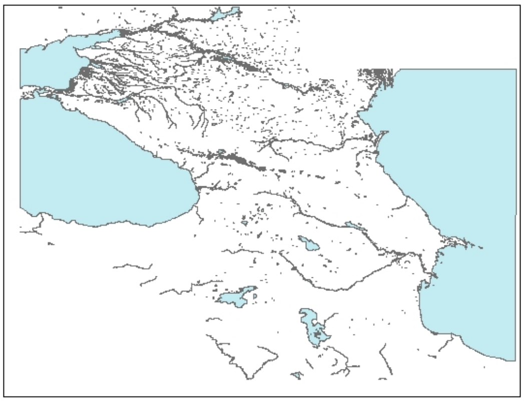
Ethnical borders (polygons)



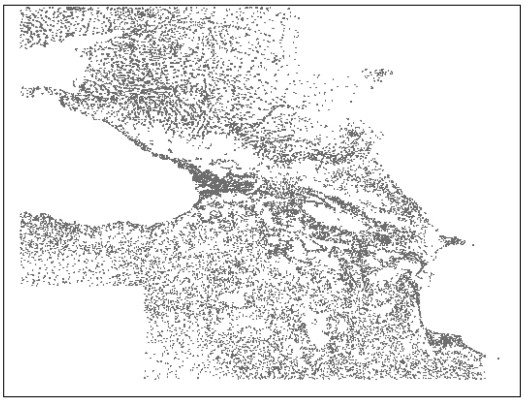
Forest (polygons)



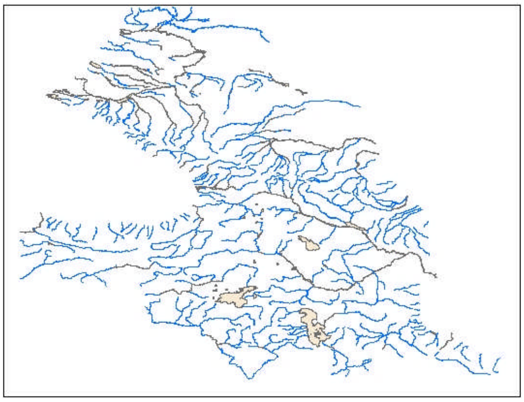
Lakes (polygons)



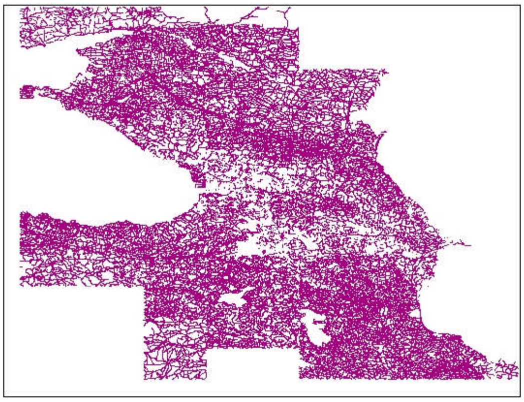
Settlements (polygons)



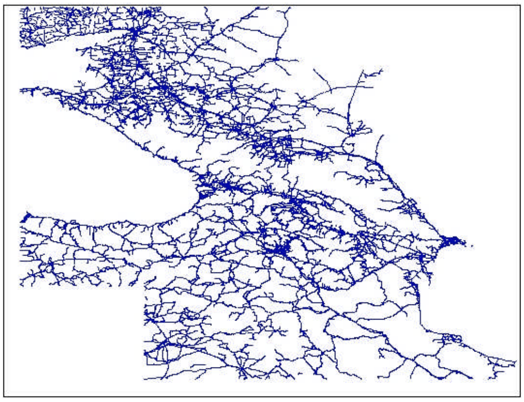
Rivers (lines)



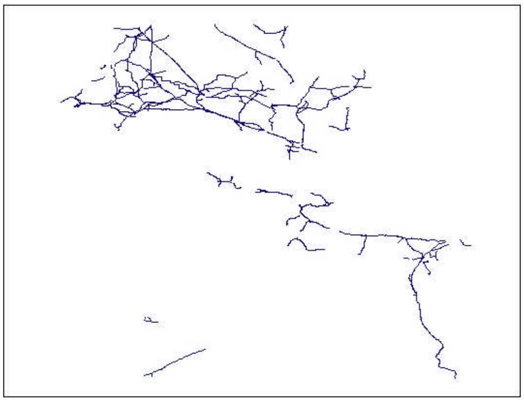
Roads (lines)



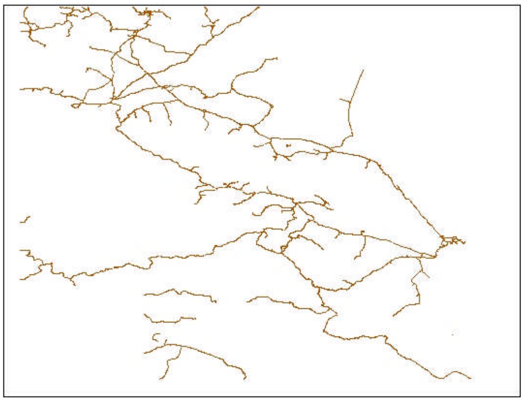
Main roads (lines)



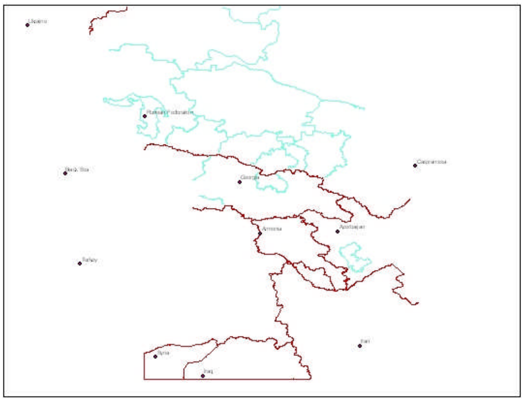
Pipelines (lines)



Railways (lines)



State and territory borders (lines)



5.1.4. Attributes information

Two types of attributes tables are associated to the entities listed above. The district and ethnic layers have two particular structures while the others entities can be grouped together because of similarities in their attribute tables.

The district layer contains data gathered at NUTS3 level in the independent countries and NUTS4 in Russia. The 158 variables refer to demographical, ethnical socio-economic and agricultural information. Table 2 lists these socio-economic parameters available for the homogenous "district" units, corresponding to different levels in the hierarchy of administrative ones. The ethnic layer refer to the main ethnic groups per polygon.

Table 2 : Socio-economic data available at district level

Statistical data at district level	Unit	Dates
Total Population	hab	1989/ 1995/ 2002/ 2005
Urban and rural population	hab	1989/ 1995 /2002
Ethnic composition	hab	1989 /2002
Birth and death rates	inhab/1000	1995 /2002
Infant death rate	inhab/1000 birth	1995 /2002
Weddings and divorce	Nbr / 1000	1995 /2002
Agicultural surfaces (cereals, vegetables, potatoes, technical cultures, vineyards, irrigated)	ha	1995 /2002
Total agricultural surface	ha	1995 /2002
Pastures	ha	1995 /2002
Forests	ha	1995 /2002
Agicultural yield (cereals, wheat, potatoes, vegetables, grapes)	Quintal/ha	1995 /2002
Agicultural production (cereals, potatoes, grapes)	Tons	2002
Agicultural production (meat, milk, wool)	Tons	1995 /2002
Cattle production (cows, sheeps and goats, pigs)	numbers	1995 /2002
Doctors	number	1995 /2002
Phones	number	2002
Cars	number	1995 /2002
Monthly wage	$	2002
Monthly pension	$	2002
Industrial production	$	2002
Total agricultural production	$	2002

The topographical entities present a similar attribute structure (See 5.1.6). Some fields are common to all the features but a lot of these fields have no meaning (Nazvanie, Saxeli, Fields 9,12,13,14). "Name" as well as "Rayun", "Status" and "Country" fields refer to the settlement layer and could identify mountainous area in the Caucasian region. Because of its incompleteness and its unclear definition, this information cannot be used in a comprehensive analysis. These attributes probably result from digitalisation or conversion mistakes. The attributes highlighted in blue in Table 3 are layers specific with partial information content. The completeness of numeric fields is estimated by the number of null values (highlighted in yellow) while the character fields are quantified by the number of blank cells (highlighted in orange).

Table 3 : Attributes of topological entities

		Lakeskav	Forestkav	MainRivKav	MainRoadKav	Poparkav	Regionline	Stateborder
Nazvanie	C	95.44%	89.34%	86.39%	83.63%	71.16%	88.24%	100.00%
Saxeli	C	91.58%	100.00%	45.38%	99.98%	40.53%	100.00%	100.00%
Name	C	90.01%	100.00%	31.06%	100.00%	0.14%	82.35%	87.50%
Typ	C	0.07%	2.12%	3.84%	0.06%	0.03%	100.00%	100.00%
Poptyp	N					100.00%		
Population	N					99.36%	100.00%	
Status	C					98.82%		
Distance	N				100.00%			
Area	N	83.85%	100.00%			100.00%	100.00%	100.00%
Lengths	N							100.00%
Length	N			13.44%				
Bassein	C	99.91%		97.03%				
Rayun	C		99.87%			89.13%		
Avtrep	C		99.28%			99.72%		
Country	C		99.39%			96.81%		
Field12	N				100.00%			
Field13	N	100.00%	100.00%	100.00%	100.00%		100.00%	
Field14	N	100.00%	100.00%	100.00%	100.00%		100.00%	100.00%
Field9	N	100.00%	100.00%	100.00%			100.00%	100.00%
		4606	3912	573	8961	24220	17	16

Legend :	% of blank cells	% of null values	C = character	N = number

Only the “typ” attribute is nearly complete and can then be used in some cases as a typology of each feature. Table 4 lists the categories provided by the producer. Unfortunately, some categories, as the “fr” and “fs” in the forest layer or “rr” and “rv” in the lakes and river layers are not defined. Provided as part of the road file in MapInfo format, railway and pipelines entities are considered as separate entities in the overall study. The mistakes in the encoding of categories detected are analysed later (See 5.4.3).

Table 4 : Entities typologies

Forest	bs	fallow
	fp	plantation
	fr, fs	forest
Lake	gl	glacier
	lk	lake
	lp	non permanent lake or lagoon
	rr, rv	river
	rv	river
	rz	artificial lake
	sea	sea
	zl	bay
Roads	as	motorway
	ru	building road
	rw	railway
	ss	road
	su	state road
	gg	path
	gu	good path

	pd	forest path
	gp	gas pipe
	ep	oil pipe
	ul	urban road
	vl	village road
Settlement	cc	capital
	ct	center of region
	kc	center of district
	rc	large town
	tw	town
	vl	villages

5.1.5. Sources of observation

The attribute statistical dataset attached to the district level is compiled through local contacts with national statistical committees and manual encoding of these data. The socio-economic data include ethnic, linguistic and religious distribution, income and agricultural information (Table 2).

The supposed source of topographical information should be the Soviet military topographic maps at 200k resolution. Several scales of these maps are available. These topographical maps are available on paper or scan format. Students and professors pertaining to the regional research team have then digitised, without protocol or technical instructions, part of information reproduced on Soviet maps (roads, railways, energy corridors as well as rivers, settlements and place names).

5.1.6. Compilation methods and internal relationships between entities

MapInfo is the software chosen to compile these data because of its availability in local institutions. This dataset has been developed during a time period of 20 years. This long term process means that changes in hardware and software versions cause heterogeneities in dataset structure. Because of various producers, sources of information, old fashioned technical education in digitalization techniques, the accuracy and consistency of data are especially low. Visual examination of parentage, redundancies and gaps between dataset state an overal low level of correctness. Overlay of layers doesn’t match perfectly at a scale lower than 1:1 000 000. For example, state borders do coincide with districts boundaries.

Taking into account these issues, the use of a GIS software to compile the data gives also a basic spatial homogeneity of the different entities. Moreover, as these data have been gathered to provide a geographical illustration of regional geopolitical changes, this objective matches to the overall goal of our instability analysis. The entities integrated in this mapping exercise correspond to the major geopolitical factors of the Caucasus complex and gives an regional picture of geographical and political contrasts and socio-economic context with a common spatial reference.

5.1.7. Transformations in the process of developing the data set

For the producer objective that is creating maps for a geopolitical atlas, data are presented in Adobe Illustrator format. Errors are corrected in the last step of the producing chain and not in the original files. Delivered in MapInfo format, the dataset has been converted to ESRI-shape-format because of user requirement. Miscoding in the dataset can be related to this conversion. Meanwhile, this conversion reveals also some digitalization or encoding mistakes. In both cases, these errors have to be identified and deleted when it is possible.

Table 5 : Types of geometry respectively in ArcGis and MapInfo software

ArcGIS	MapInfo
Point	Point
Line	Line (Single line with no nodes)
	Polyline (Line with nodes)
	Arc
Polygon	Polygon
	Region
	Ellipse
	Rectangle
	Rounded Rectangle

As the internal data management of MapInfo software differs from ArcGIS a single MapInfo map file can contain many different types of geometry (point, line, polyline, arc, ellipse, rectangle, rounded rectangle, region, and text) (Table 5). The MapInfo format also stores features with no geometry. Features having no geometry are referred to *none* geometry. Table 6 summarize the different shape-files converted from the original MapInfo coverage and the ones selected as relevant.

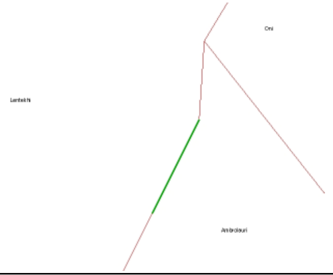
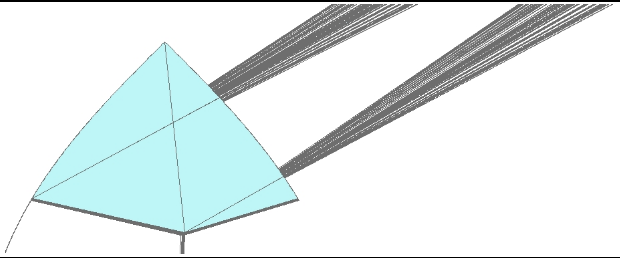
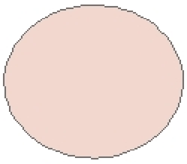
Table 6 : Transfer of MapInfo files in ArcGis format

Themes	Shapefiles created in the transfer-name (*.shp)	Feature type	Number of objects	Shapefile selected for consolidation
Districts	DistrictISPRA_polyline	Polyline	1	
	DistrictISPRA_region	Polygon	357	DistrictISPRA_region
State border	state_border_polyline	Polyline	16	state_border_polyline
	state_border_text	Point	11	
	region_line_polyline	Polyline	17	region_line_polyline
settlement	Poparkav_arc	Polyline	1	
	Poparkav_none	Point	18	
	Poparkav_point	Point	254	
	Poparkav_rectangle	Polygon	10	
	Poparkav_region	Polygon	24220	Poparkav_region
	Poparkav_rounded_rectangle	Polygon	65	
Roads, pipelines and railways	RoadKav_none	Null	3248	
	RoadKav_point	Point	1	
	RoadKav_polyline	Polyline	34006	
	RoadKav_text	Point	1	
	MainRoadKav_polyline	Polyline	8961	MainRoadKav
	MainRoadKav_point	Point	1	
	MainRoadKav_none	Null	702	Extraction of :
	MainRoadNF_none	Null	351	Pipelines
	MainRoadNF_polyline	Polyline	3768	Railways
	MainRoadNF_point	Point	1	
	MainRoadNF_none	Point	702 (small extent)	
Rivers	MainRivKav_polyline	Polyline	573	MainRivKav_polyline
	MainRivKav_region	Polygon	294	MainRivKav_region
Lakes	Lakeskav_region	Polygon	4606	Lakeskav_region
	Lakeskav_polyline	Polyline	3 (redundant with river)	
	Lakeskav_none	Null	1-shape null	
	Lakeskav_ellipse	Polygon	8	
Forest	Forestkav_arc	Polyline	1	Forestkav_region
	Forestkav_ellipse	Polygon	3	
	Forestkav_none	Null	7-shape null	

	Forestkav_polyline Forestkav_region	Polyline Polygon	465 3912	

Based on information provided by the producer and visual examination of the resulting files, some shapefiles have been rejected – non relevant- (Table 7). These files present no interest either for the producer or the user. The irrelevance of these features can usually be explained by bad digitalization process. Topological and logical rules allow to discard these files because of, for example, disjoined features, crossing lines or different precision in the digitalization on some part of the region. The transfer process is then seen as a way to identify these features. Table 7 illustrates some irrelevant shape files.

Table 7 : Examples of typical errors within the topographical shape files

DistrictISPRA_polyline	1	
Poparkav_arc	1	
Poparkav_rounded_rectangle	65	
Lakeskav_ellipse	8	

5.1.8. Spatial reference of the original dataset

Two main coordinate systems are used in the original data. The district layer providing the reference statistical units uses a specific projected coordinate system (Afgooye / UTM zone 38N) and the other dataset have a lat long geographic coordinate system.

- *Projected coordinate system*

Projection: Transverse_Mercator
False_Easting: 500000.000000
False_Northing: 0.000000
Central_Meridian: 45.000000
Scale_Factor: 0.999600
Latitude_Of_Origin: 0.000000
Linear Unit: Meter (1.000000)

Geographic Coordinate System GCS_Afgooye
Angular Unit: Degree (0.0174532925199433295)
Prime Meridian: Greenwich (0.000000000000000000)
Datum: D_Afgooye
Spheroid: Krasovsky_1940
Semimajor Axis: 6378245.000000000000000000
Semiminor Axis: 6356863.018773047300000000
Inverse Flattening: 298.300000000000010000

- *Geographic Coordinate System :* Lat Long for MAPINFO type 0 Datum
  Geographic Coordinate Units: Decimal degrees (Angular) (0.0174532925199433299)
  Prime Meridian: Greenwich
  Latitude Resolution: 0.000000
  Longitude Resolution: 0.000000
  Geodetic Model
  Horizontal Datum Name: D_MAPINFO
  Ellipsoid Name: World_Geodetic_System_of_1984
  Semi-major Axis: 6378137.000000
  Denominator of Flattening Ratio: 298.257224

5.2. Scope of the represented real world

As already mentioned, these data have been acquired because of their unique character of depicting the major geopolitical issues in the transnational region of Caucasus. Gathered to provide a geographical illustration of regional geopolitical changes, the resulting data set match to the overall goal of our instability analysis.

5.2.1. For the geopolitical atlas (producer)

The Geopolitical Atlas project aims to develop an analytic tool as well as a way to **represent the physical, spatial and socio-economic attributes of the Caucasus**. This effort implies a multi-disciplinary analysis comparing underlining regional and local specificities and inter-twining multiple and evolving contributions of geography, history, political science, sociology and economy. The project shed light on the redistribution of demographic, economic and political long term phenomena as key elements for the understanding present situations and factors of change.

Table 8 : Multi-disciplinary factors of Caucasian geopolitics

History	Different definition of caucasian territories
	"Dreamed" or historical territories
	Evolution of the administrative units
History of conflicts	Caucasian conflicts since 1988
	Territorial and border contests
Population and Demography	Population density by district (hab/km2)
	Urban Population (%) and city size
	Population growth (189/2002/2005)
	Birth ; Death ; Natural growth by district
	Infant mortality
	Migration, regional
The ethnic mosaic	Ethno-linguistic Map of the Caucasus
	Religions
	First Nationality -nation 1-, by district 2002

	Second nationality -nation 2-, by district 2002
	Eponym population by district 1989/ 2002
	Russians population evolution
Economy	PIB per capita and evolution per region
	Active population and unemployment per region
	Main economic projects
	Electricity (production and consumption per region)
	Industry per capita
	Main industrial plants
	Agricultural production per capita, by district
	Wood and wood processing
Social development and disparities	Poverty
	Population equipment (phones and cars)
	Salary / pension
	Health sector equipment
	Crime
Transport and foreign trade	TRACECA; Tubes and oil transport
	The ways of traffic (air and airports)
	Foreign trade by States

At the same time, the project attempts to confirm the relevance of the proposed limitation of the Caucasus territory, be it administrative, natural, historical or new geopolitical. The political frontier between the Northern Caucasus which belongs to the Russian Federation and the independent states of Southern Caucasus is one of the obstacles to a global approach of this historical and geographical unit.

This project tries to verify the relevance of such an approach by creating a large data bank of new processes concerning the Caucasus and, by using analysis tools of cartography integrated in a regional atlas. The preparation of such an atlas implies the pre-elaboration of a multi-factorial data base, collecting a set of geographic, demographic, ethnic, socio-economic, historical and cultural data at different scales. The four population censuses that took place within the last two years in Russia as well as in the three southern Caucasian republics offer a splendid opportunity for an overview with renewed data. These censuses, even with their methodological bias, will appropriately complete the data base necessary for this study. Maps are used as a fundamental analysis tool at the district scale that highlight the complexity of ethnic and social phenomena in such a heterogeneous zone.

5.2.2. For the geomodel of instability (user)

Defining by a formal statement the universe intended to be represented by the data is particularly challenging in an instability/insecurity analysis. “To be secure is to feel free from threats, anxiety or danger” (Art 1993). Using this definition, security is a state of the mind of the people, a feeling, a perception varying for each individual and for each level of political decision from local to international. So there is an infinity of security definitions. Instability is context dependent (space and time). Geography matters. In our analysis we choose to address **the regional territorial security**. This notion is related to geopolitical security issues as territorial contiguity, territorial belonging, sovereignty within borders that are essentially contested notions in security research (Stephenne and Ehrlich, forthcoming).

Classical security complex theory posits the existence of regional subsystems as objects of security analysis and offer a analytical framework for dealing with those systems. All the states in the system are integrated in a global web of security interdependence. But because most political and military threats travel more easily over short distance, insecurity is often associated with Proximity. Most states fear their neighbour more than distant powers, security interdependence across the international system as a whole is far from uniform (Buzan et al. 1998). The coverage of our analysis is restricted to the

Caucasus regional complex : the three independent states and the Caucasian regions in Russian Federation. The Caucasus instability cannot be assessed without taking into account all the transboundary issues of this territory.

A model of instability/insecurity means looking at the contextual " predisposing causes of threats". There are different kind of freedoms (political, economical, social) interlinked to each other (Sen 1999). The predisposing causes of insecurity, already defined by Snow in 1855, are the characteristics of person and places that determine the impact of a given threat (quoted by Webb and Harinarayan 1999). Our instability study is based on the idea that a better understanding of complex interactions between threats affecting human security can be addressed through a modular modeling approach.

Table 9 : Expanded concepts of Security (adapted from Brauch 2005)

	Reference object (security of whom?)	Value at risk (security of what?)	Source(s) of threat (security from whom or what?)
National Security (political, military dimension)	The State	Sovereignty, territorial integrity	Other states, terrorism (substate actors)
Societal / territorial security	Nations, societal groups	National unity (borders, distribution of resources), Identity (population distribution)	Nations, migrants, aliens Cultures religions (States)
Environmental security	Ecosystem	Sustainability	Humankind
Human security	Individuals humankind	Survival, quality of life	State, globalisation, nature, terrorism
Gender security	Gender relations, indigenous people, minorities	Equality, identity, solidarity	Patriarchy, totalitarian institutions (governments, religions, elites, culture), intolerance

5.3. Control reference data

Quality assessment of geospatial data is an exercise of relative performance. Performance of an assessment requires comparison of the test data against some reference. By definition, the data set is a representation of real world phenomena and then a simplification of these phenomena. The evaluation of this representation can only be accomplished by comparing the result against the intended model. As the scope of our intended model is an abstraction, it is kept as a background objective but a source of higher quality data typically serves as a comparative model for positional accuracy tests. A data set is considered of higher quality than the test data if it has one or more of the following characteristics:

- Represents more detail (is at a larger scale).
- More rigorous data quality assurance procedures were known to be used in the data collection.
- Made use of instrumentation known to be of higher quality.
- Comprises a more recent measurement.
- Consists of direct observations/measurements in the field.

No data set provided complete coverage of the study area for all the Caucasian information layers, especially with superior accuracy to those being assessed. Data quality assessment therefore has to rely on "best available" data for each dataset layer. The control data sets can be categorised due to their link with the data source (Table 10). This categorisation will help us to describe the different control dataset proposed in this assessment.

Table 10 :Categories of control data

Primary control data	Data accuracy is known, data has a detailed linage and complete metadata.
Secondary control data	Data created by digitizing from primary control data. Digitized

	features allow for vector on vector analysis to be done if the primary control data is in raster format.
Tertiary control data	Data accuracy is not know but shows good or spatial similarity or compares well at a specified scale, when compared with primary control data.

Table 11 lists and details the primary data sets. For most of the layer, the Soviet topographical maps are the best data comparison source, with at least two of the criteria defined before : (i) a larger scale (100k topographic maps is used as control dataset while the original source is supposed to be the 200K) and (ii) a rigorous quality assurance associated to all the Russian cartographic product. Moreover, the link of dataset to the supposed original source of the dataset (See Section 5.1.5) is an interesting component to assess the quality of the original digitalisation.

Following the categorisation proposed, the 100k Soviet topographical map is a primary control data that provides the best available accuracy. However, two major problems persist. Firstly this data is in raster format which requires digitising if comparative analysis of vector data is to be done. Digitalization requires a high amount of time. This makes difficult to test it on large area. Secondly, these data were not available on the full study area.

Table 11 : Available primary control data sets

Name	Data Type	Content	Scale/resolution	Coverage	Detail
Soviet Maps 100k	Raster	Topographical	1: 100 000	Partial availability	Excellent
Soviet Maps 200k	Raster	Topographical	1: 200 000	Partial availability	Good
TPC maps	Raster	Topographical	1: 500 000	Partial	Medium
Quickbird Image	Raster	Satellite image	60 – 70cm pixel size	Partial	Excellent
Landsat Image	Raster	Satellite image	15m pixel size	Complete	Excellent

Mostly because of time and resource constraints, the accuracy assessment cannot be carried out with the same quality and level of details for all the layers. The digitalisation of the Russian maps requires a lot of time combined with high level of technical GIS education. The illustrative protocol that is proposed uses the “settlement layer” to set up the accuracy assessment methodology without applying it to all layers. Because its resolution and precision presents the highest accuracy, the assessment of this layer accuracy should provide an overall indicator about the quality of the overall dataset. For the other layers, this indicative quality value is completed by an assessment based on tertiary control data.

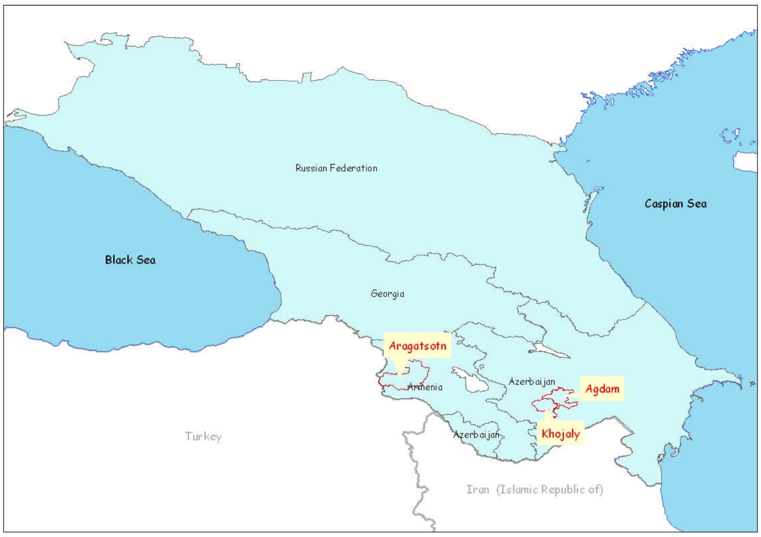
5.3.1. Test areas

Data comparisons must be made in the same geographical area, for this reason it is important to have a clear definition of the study area so irrelevant data may not be included into spatial and statistical analysis. The study area also provides a means of assessing the completeness of the involved data by the comparative geographical coverage of the data and the study area. As mentioned before, the Caucasian study area is defined by the 363 districts. This area includes Armenia, Azerbaijan, Georgia and a southern part of the Russian federation. The study area is boarded by the Caspian Sea in the east and the Black sea in the west.

For evaluating the accuracy of some data layers, two test areas were chosen based on district borders (i.e, Aragatsotn in Armenia, Agdam and Khojaly in Nagorno-Karabagh). Unfortunately, two test areas are not sufficient to be representative of the physical and geographical character of the region. No extrapolation of results can then be foreseen. But as the accuracy assessment is time consuming, it has not been possible to increase the number of test areas and these results are illustratives. Figure 2 shows the location of these areas in relation to the overall study area. These areas (2731km^2 for Aragatsotn, 1135 km^2 for Agdam, and 936km^2 for Khojaly) only represent 1.1% of the spatial coverage of the overall Caucasian districts (436 785 km^2). The selection of these test areas is based on 5 criteria:

1. one district of the study area
2. representing as much as possible some diversity of the regional instability (choice of one conflicting area- NK / and one test area that is not in a conflicting part of the region)
3. primary control data available
4. existing data from as various sources as possible
5. areas of similar size.

Figure 2 : Districts selected for the specific topographical test area



5.3.2. Comparison control data availability

The key component for ensuring the reliability of the accuracy tests is the selection of a source of higher accuracy. Because of the lack of information about the origins and the processing of the Caucasus dataset but also because of the various entities covering a large study area, a mixed and adapted strategy has been set up. Existing control data of various origins and various resolutions can been used in combination to proper digitalised features (Table 12). Because of time constraints, these control data have only been partially integrated in the geodatabase accuracy test until now. Remote Sensing data in particular, can be further involved in accuracy evaluation or data consolidation at a more local scale.

Table 12 : Control data available

Feature	Comparative/ Control data			Best scale / resolution	Test area / coverage
	Primary	Secondary	Tertiary		
District Boundaries	Gaul data			1:1 000 000	Global
	Landsat 7			15m pixel size	Global
Settlements		Digitalisation from Soviet 100K		1: 100 000	Agdam, Aragatsotn
			polygons		Aragatsotn
			points		Khojaly
	Quickbird			0,7m pixel size	Agdam
Roads	AZE-Nima		Road lines	1:1 000 000	Aragatsotn, Agdam, Khojaly
Pipes	Energy Map		Pipe lines	Not localised	Regional
Rivers	Geocom, ltd		River lines	1:200 000	Armenia
Forest	Landsat 7/ Quickbird			0.7, 15m pixel size	Agdam

5.3.3. Primary control data sources description

5.3.3.1.Soviet maps

For the 50 years prior to the collapse of the Soviet Union in the early 1990s, the Soviet military sought to map every corner of the globe. The result was an extensive collection of standardized maps at various scales. Since its formal establishment in 1812, the soviet topographical military mapping project is leaded by the Chief Administration of Geodesy and Cartography (GUGK), principal topographic map producing organization of the formal USSR. The Military Topographic Administration (VTU) under the General Staff of the Ministry of Defense is also involved in Soviet mapping, operating in close cooperation with GUGK. Its responsibilities are flexible but apparently its authority takes precedence during wartime. During the Second World War, the need for large scale mapping of European Russia was particularly acute and was met with production of 13,000 map sheets printed in the first six months of the war.

It is estimated that the mapping program produced over 1 million separate sheets, 800,000 for the USSR alone (Davies John, 2006, Sheetlines from Charles Close Society) In his paper, Davies indicates that the Soviets mapped the entire world at 1:1,000,000, 1:500,000 and 1:200,000, most of Asia, Europe, north Africa and North America at 1:100,000, the Soviet Union, Europe and parts of Asia at 1:50,000, the Soviet Union and eastern Europe at 1:25,000 and about a quarter of the Soviet Union at 1:10,000. “In addition,” writes Davies, “plans at 1:25,000 and 1:10,000 were produced of thousands of towns and cities around the world.” In some areas, the Soviet maps are still among the best available.

5.3.3.2.GAUL

The basic district layer combines administrative units at different level of details within the four countries. At the level of the state, boundaries are not only imprecisely localised but more often the border and territorial unit is not agreed between the two countries. Consequently to these political reasons, border lines in the district layer represent a quite international agreement rather than the position of one of the conflicting positions. The international project compiling the country and administrative units boundaries is called Global Administrative Unit Layers (GAUL[2]).

GAUL is an initiative implemented by FAO within the EC-FAO Food Security Programme funded by the European Commission. The GAUL aims at compiling and disseminating the most reliable spatial information on administrative units for all the countries in the world, providing a contribution to the standardization of the spatial dataset representing administrative units. The GAUL always maintains global layers with a unified coding system at country, first (e.g. regions) and second administrative levels (e.g. districts called “Gaul Admin 2 boundaries”). In addition, when data is available, it provides layers on a country by country basis down to third, fourth and lowers levels.

The overall GAUL methodology consists in a) collecting the best available data from most reliable sources, b) establishing validation periods of the geographic features (when possible), c) adding selected data to the global layer based on the country boundaries provided by the UN Cartographic Unit version 5 (UNCS), d) generating codes using the GAUL Coding System and e) distribute data to the users.

We used this dataset to (a) check the accuracy of the boundaries, (b) identify the discrepancies between the two sources and (c) import the standardized codes in the district dataset.

5.3.3.3.Satellite imagery (Landsat and VHR data)

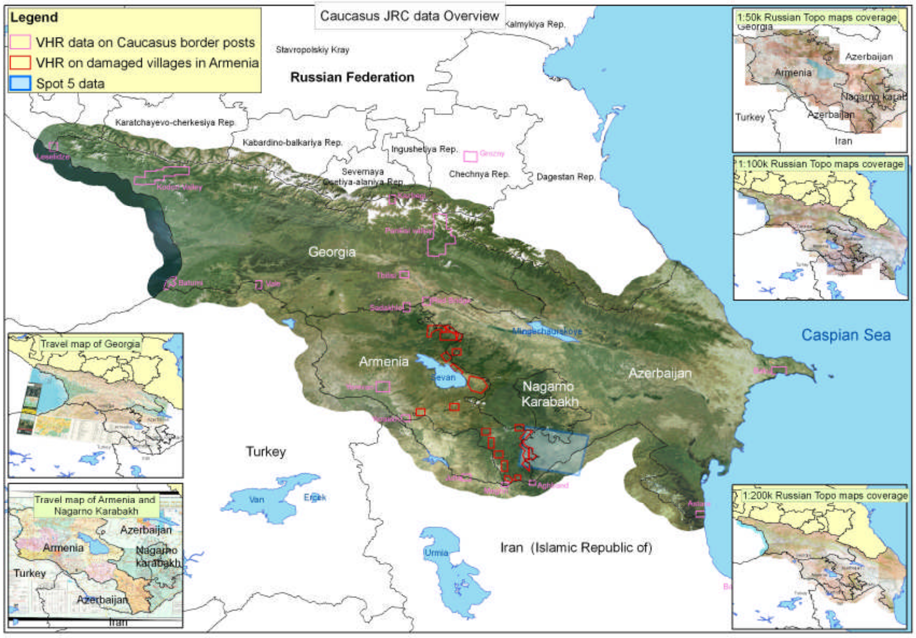
Landsat ETM imagery is available globally for the year 2000 (GeoCover Landsat mosaics) and provides a comprehensive coverage of the Caucasian region. The low resolution of these data (pixel size of 28.5 meters) is not sufficient to identify the entities presented in our dataset, especially the

[2] http://www.foodsec.org/News/tools_gaul.htm

settlement feature. Because of digital processing and global use, these data are provided with a controlled absolute positional accuracy of 50 meters (Root Mean Square Error). This positional accuracy and the true colour representation of the landscape can be useful in a visual comparison with the Caucasian dataset.

With a much higher resolution (Very High Resolution – VHR), two other sensors take satellite images: Quickbird (0.61m in panchromatic mode) and Ikonos (1m in panchromatic mode). Because of the cost of these data, they are not available on the whole Caucasus, but can be used as a complementary comparison on some test areas. These images could potentially be used as inputs for the digitalisation of control data in a precise assessment at local scale.

Figure 3 : Remote sensing data (Landsat coverage and VHR) for visual comparison or digitalisation input

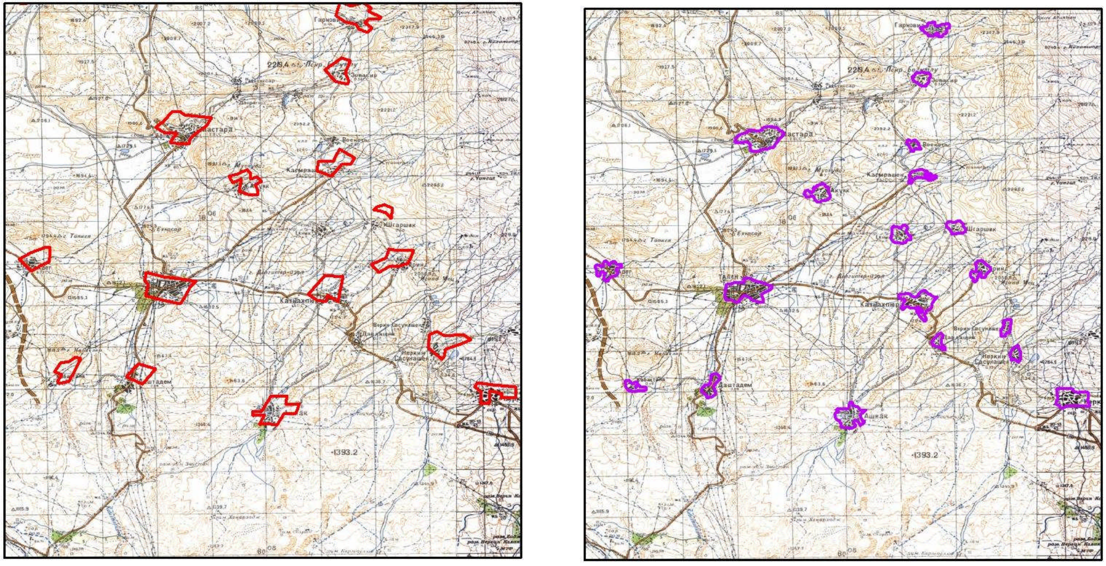

Legend
VHR data on Caucasus border posts
VHR on damaged villages in Armenia
Spot 5 data
Caucasus JRC data Overview
Kalmykiya Rep.
Stavropolskiy Kray
Russian Federation
Karatchayevo-cherkesiya Rep.
Kabardino-balkariya Rep.
Ingushetiya Rep.
Grozny
Severnaya Osetiya-alaniya Rep.
Chechnya Rep.
Dagestan Rep.
Kodori Valley
Georgia
Batumi
Tbilisi
Sadakhlo
Mingechaurskoye
Armenia
Sevan
Nagarno Karabakh
Azerbaijan
Caspian Sea
Turkey
Van
Ercek
Urmia
Iran (Islamic Republic of)
1:50k Russian Topo maps coverage
1:100k Russian Topo maps coverage
1:200k Russian Topo maps coverage
Travel map of Georgia
Travel map of Armenia and Nagarno Karabakh
Iran

5.3.4. Digitalised dataset

Digitizing can provide a time efficient method for control data collection. General sources for developing digital data include paper maps, aerial photos, digital orthophoto and satellite imagery. Two type of digitalised dataset can be distinguish in this assessment: (i) the secondary source, referring to polygons and lines directly digitalised for the purpose of this study but only for the settlement information, and (ii) the tertiary source referring to dataset acquired from other sources with unknown quality.

For the intended purpose of the settlement entity assessment, the highest quality source is the topographical soviet map at 100K. The topological soviet map at lowest scale (200K) is the supposed source of the overall dataset. The use of same source, with common hierarchical methodology, but with highest resolution creates the 100% assurance that the digital control source is of higher quality. The fact that the digitalisation process has been carried out by one individual on both test areas provides the homogeneity of the resulting control data.

Figure 4 : Caucasian (test) and digitalized (control) settlement dataset in Aragatsotn district on 100k topological map.



Some tertiary sources of higher accuracy are used in this assessment: (i) a database including rivers, lakes, railways, roads and settlements for the entire Armenia extension, (ii) a less complete but more precise database including only rivers, roads and points for the settlements on the Nagorno-Karabakh region, (iii) a main road dataset (Aze-Nima global dataset), and (iv) the Energy Map of the Middle East and Caspian Sea Areas (Petroleum Economist Ltd.). The two first ones have been produced by GIS consultant company (Geocom, ltd.). Unfortunately, these data have no metadata but present a better visual matching with Soviet topological maps. The source of the data with Armenian coverage are probably the 200K topological maps while the Nagorno-Karabakh data have a better precision with more objects and then could refer to the 100k. The last road dataset is extracted from the US National Imagery and Mapping Agency (NIMA) global dataset and is used below in a general assessment of main roads positional accuracy.

Figure 5 : Tertiary control datasets for the river entity available on Aragatson and Khojaly test areas

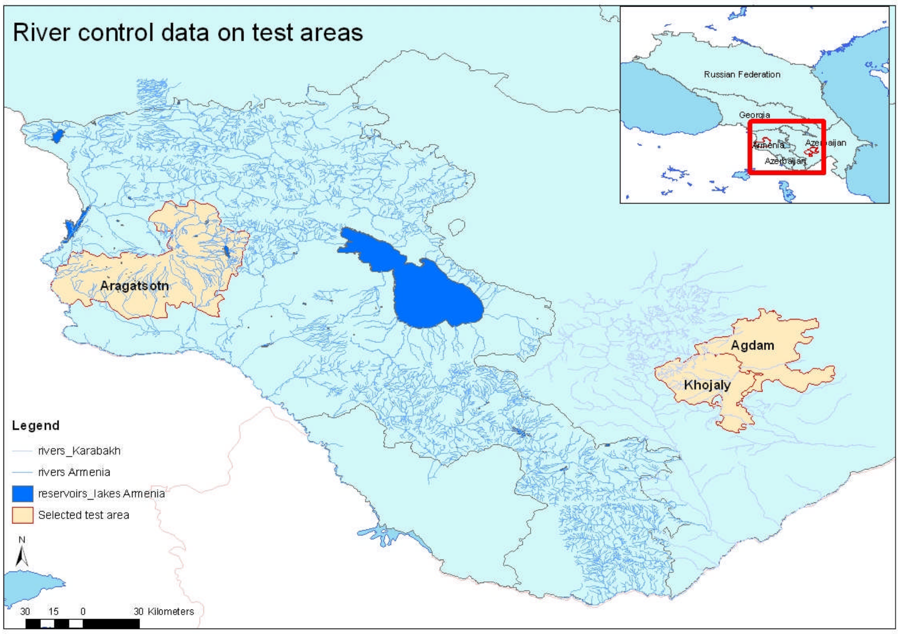
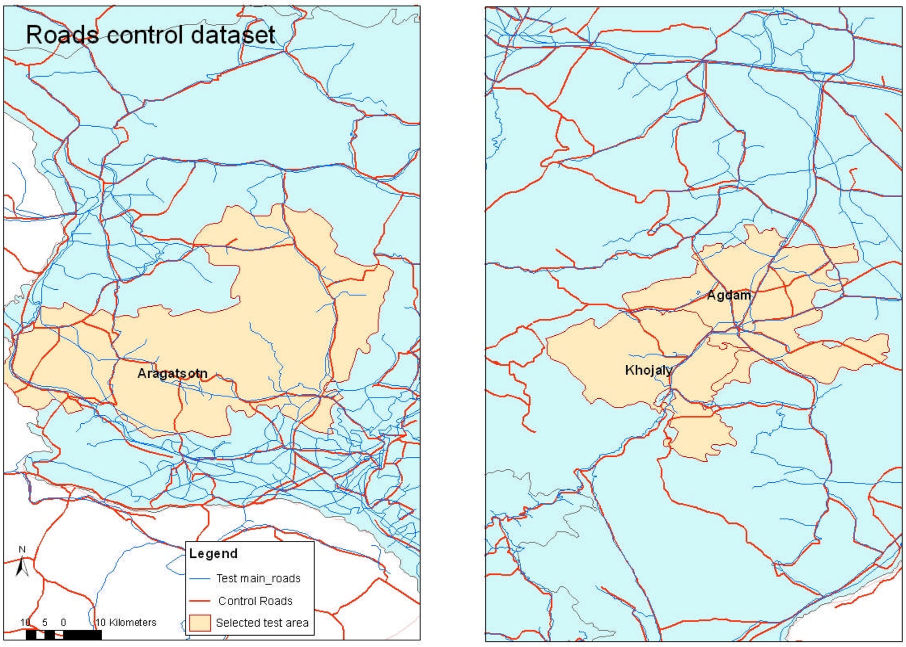


Figure 6 : Control dataset proposed to asses the main roads position in the three test areas



5.4. Accuracy

Three data accuracy dimensions (positional, temporal and thematic) are assessed in this section. As stated in the previous section, a mosaic of evaluation methods are put together in order to quantify the overall accuracy of the database. The comparison of methods allow to cross-check the different results and provide a better picture of the reliability of these data. The majority of the assessment effort focused then on positional accuracy and settlement layer. As explained before (See 5.1.4), the assessment of the trueness of topographical attributes categories refers mainly to the typology, while Table 3 illustrated the completeness of the others attributes.

5.4.1. Positional accuracy

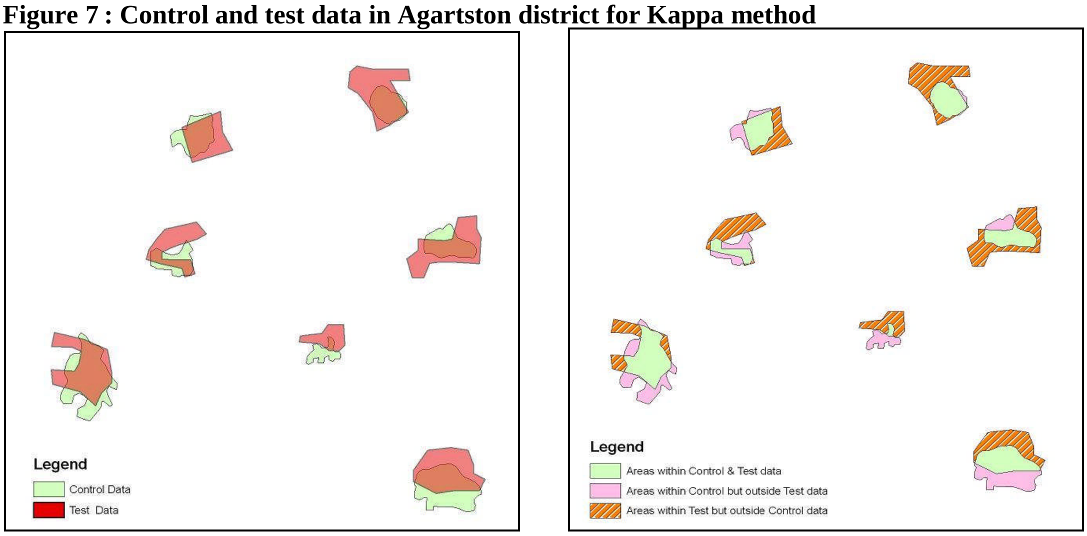
Depending on the type of entity, point or linear features, the assessment method differ. For point entities, the positional accuracy is measured by the standardized method of the root mean square error (RMSE) (Spatial Data Accuracy handbook, 1998). RMSE is the square root of the average of the set of differences between the coordinates values for the test data and the control data. Positional accuracy is measured independently in the horizontal and vertical directions. The task that is most critical in ensuring the reliability of this test is the correct matching of the control and test points. The procedure for assessing horizontal positional accuracy consists of the following steps:

1. Collect x and y position measurements for the point objects in the control and test data sets.
2. Match the control points to the appropriate points in the test data set, with a spatial join.
3. For the matched points, calculate the radial root mean square error (RMSE): (S ((control x – test x)2 + (control y – test y)2)) / number of matched points)1/2
4. Adjust the RMSE for a 95% confidence interval: RMSE * 1.7308

For polygons entities, An alternative approach to representing polygon error is a correlation statistic called Kappa, presented by Greenland, Socher, and Thompson (1985). Kappa calculates the percent correctness of a map and allows for comparison to other maps (Congalton, 1991). Studies have found it useful and credible for analyzing the relative strengths and weaknesses of two data sets (Greenland et al, 1985). An example illustrates the methodology. In Figure 7, Part A, the solid line (C) represents the polygon object for the control data and the dashed line (T) represents the polygon object for the test data. The result of an overlay procedure is displayed in Figure 7, Part B. Four distinct classifications of areas are derived from the overlay operation:

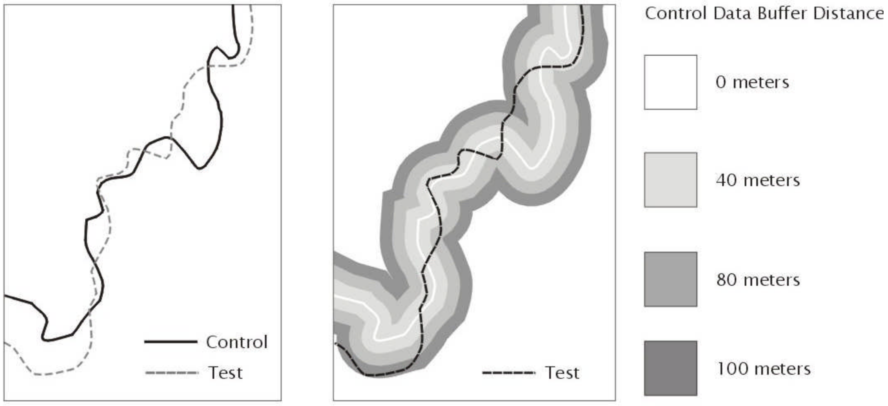
- Areas located within both Control and Test
- Areas located within Control, but outside Test
- Areas located within Test, but outside of Control
- Areas located outside both Control and Test

Figure 7 : Control and test data in Agartston district for Kappa method


Legend
🟩 Control Data
🟥 Test Data
Legend
🟩 Areas within Control & Test data
🟪 Areas within Control but outside Test data
▨ Areas within Test but outside Control data

For linear entities, a non-parametric alternative method to the point measurement is the distance buffering method (Goodchild and Hunter 1997). These authors consider a buffer of width x around the reference source and compute the proportion of the tested source length that lies within the buffer (Figure 8). This approach provides a percentile distribution of accuracy and could be generalized to area features.

Figure 8 : Distance buffer method for positional assessment of linear features (from Goodchild and Hunter 1997)



A. Overlay of Control and Test Data of a Stream Segment

B. Process of Clipping the Test Data From a Buffer Distance

Table 13 summarizes methods applied to datasets and relevant control data. As clearly stated in this table, every comparison differ either in the control data used or in the type of method. These assessments allows to relatively quantify the quality of the data knowing the low but assessed quality of control data. These localized data presenting a higher positional precision than our dataset provide some guidelines and reference in the future use of the Caucasian dataset.

Table 13 : Mosaic of positional assessment methods

Data	Control data	Test Areas	No. of test points	No. of line test segments	Method
Ethnic	No				
District	Gaul Admin 2 boundaries	Entire dataset		300	Buffer
Settlements	Re-digitized (200k soviet maps)	Agdam, Aragatsotn, Khojaly	20		RMSE
Lakes	Quickbird	Agdam			Visual
Forest		Agdam			Visual
Rivers	Rivers (tertiary)	Agdam, Aragatsotn, Khojaly			Buffer/
Railway					
Pipes	Middle East	Region			Visual
Main Roads	AZE_NIMA Main roads	Agdam, Aragatsotn, Khojaly		93	Buffer/ RMSE

For the settlement dataset, point and polygon methods have been used in order to compare the results and evaluate the interest of one or the other in our specific case. RMSE method applied to the three test areas measures an overall accuracy of 423m (Table 14). Kappa method has been applied only on the settlements polygons of Aragatsotn with a result of 64% of agreement and 36% of Kappa. Of course this result is really low but still quite realistic knowing the lineage of the data. This quantitative assessment provides an interesting objective value. The small area used in this test comparing to the overall region has to be kept in mind.

Table 14 : RMSE settlement assessment in the three test areas

Residuals [Meters]		Residuals Squared	
Control(X)-X = V(X)	Control (Y)-Y = V(Y)	V(X)*2	V(Y)*2
-292.433	-132.9153	85516.90	17666.47
38.089	247.1395	1450.78	61077.93
90.300	391.3738	8154.14	153173.46
13.253	98.3365	175.63	9670.07
61.604	174.6376	3795.05	30498.31
208.364	99.4816	43415.37	9896.58
-28.794	-50.0316	829.07	2503.16
174.73	311.95	30532.01	97310.38
328.66	360.17	108018.34	129720.32
89.63	343.69	8034.21	118124.80
278.10	295.12	77341.30	87097.98
39.44	-54.19	1555.47	2936.31
-67.28	347.37	4526.03	120666.13
-571.52	463.54	326632.76	214865.98
342.71	643.54	117451.61	414141.00
-228.25	330.79	52096.18	109424.41
-388.11	493.66	150630.12	243700.41
335.20	374.25	112359.22	140061.24
143.59	159.53	20618.48	25449.90
205.63	277.09	42282.17	76778.76
-120.87	703.25	14609.01	494558.42
Number of samples:		21	21
Sum of Residuals squared:		1210024	2559322

RMSE of each coordinate:	240.0	349.1
Circular RMSE (X,Y):	423.7	
Meter horizontal accuracy at 95% confidence level.		

Table 15 : Kappa test for settlements - District of Aragatsotn

		Area (m2)		
Areas within Control & Test data		34659428.06		
Areas within Control but outside Test data		32222986.59		
Areas within Test but outside Control data		36773810.45		
Areas Outside Test and Control data		0		
Total area		103656225.1		
		Classified by Test data		
		IN	OUT	
Classified by Control data	IN	0.33436900	0.31086398	0.64523298
	OUT	0.35476702	0.00000000	0.35476702
		0.68913602	0.31086398	1.00000000

Percentage of agreement	64.52%
Expected fraction of Agreement	44.31%
Kappa Statistics	36.29%

For the linear entities as the districts (Table 16), the roads (Table 17) and the rivers, the buffer method give the following results at least 50% of the lines are included in the buffer of 500m but 4000m are necessary to enclose 90% of the lines.

Table 16 : Buffer/Clip Results on district /Gaul Data

Buffer Size (meters)	Frequency	Sum of lengths (meters)	% of line Within Buffer	Cumulative % of line Within Buffer
200	1125	14994556	33	33%
500	989	7111105.841	16	49%
1000	905	8132625.988	18	66%
2000	780	6887969.901	15	81%
3000	523	2988337.357	7	88%
4000	419	1679375.102	4	92%
5000	349	1376144.078	3	95%
6000	264	756448.3582	2	96%
10000	219	1630707.998	4	100%

Table 17 : Main roads/ AZE_NIMA main roads (secondary control data) : Results from the Buffer/Clip Process

Buffer Size (meters)	Frequency	Sum of lengths (meters)	% of line Within Buffer	Cumulative % of line Within Buffer
200	64	156589.2587	31	32%
500	84	162072.1482	33	65%
1000	48	76996.3196	16	80%
1500	32	31573.9099	6	87%
2000	22	21044.2179	4	91%
2500	20	13002.1607	3	94%
3000	15	11170.6961	2	96%
3500	13	9606.3228	2	98%
4000	13	10975.6375	2	100%

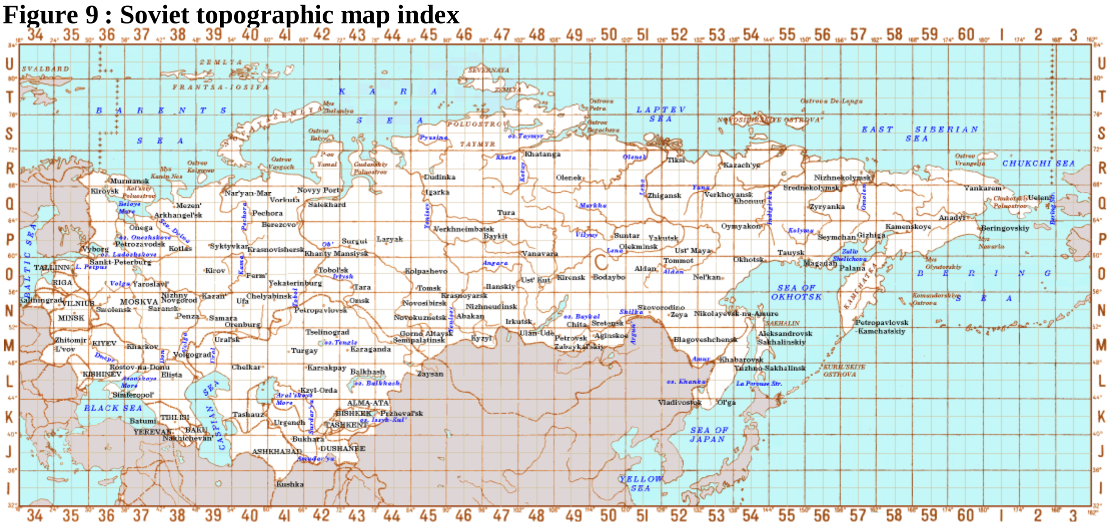
Table 18 : Armenian rivers (secondary control data) : Results from the Buffer/Clip Process

Buffer Size (meters)	Frequency	Sum of lengths (meters)	% of line Within Buffer
200	66	147979.8679	19
500	47	357138.2266	38
1000	46	106865.4279	17
1500	39	27580.5801	4
2000	35	23636.9713	4
2500	33	18836.916	3
3000	27	19188.4214	3
3500	26	16637.2493	3
4000	22	16371.217	3
4500	21	18253.5448	3
5000	19	19664.1704	3

5.4.2. Temporal accuracy

As explained in the lineage, most of the topographical layers derived from the soviet military topographical maps. The long term digitalisation phase does not allow to define either the date of the digitalisation source, or the date of the digitalisation. The production dates of the topographical maps vary between 1940 and 1990's (Davies, 2006). The Caucasian sheets - K and J 38- (Figure 9) are mainly produced between 1975 and 1985.

Figure 9 : Soviet topographic map index



Knowing the date of the original data and the geopolitical context of the region, the currentness of the dataset is highly questionable. If this dataset cannot integrates all the political updates in the different layers, the Atlas team put a lot of energy to assess these geopolitical modifications (closed roads, unavailable railways path, new or closed pipeline and closed or open borders) in different static maps (Table 8). JRC has proposed its collaboration to introduce these changing factors, illustrating the regional instability, in the geodatabase. Changes in administrative boundaries are highly frequent everywhere in the world and in particular in regions with border and territory disputes. The "stateborder" layer specifically attempts to take into account of these disputed territories and their undefined borders that are of high interest in the modelling of instability factors. One challenge of our modelling approach is to integrate this relative typology in the geodatabase (See 7).

5.4.3. Attribute accuracy

The assessment (See 5.1.4) distinguished two types of attributes referring to the statistical data associated to the district layers and the attributes fields of the so-called topographical entities. This accuracy deals with (i) district layer evaluation by the comparison with GAUL dataset, (ii) the typology information in the topographical layers and (iii) in particular a quantitative assessment on the settlement layer in the test area of Nagorno Karabakh.

The main source of positional comparison for district boundaries is the GAUL dataset already used for the quantitative evaluation of positional accuracy. The attribute and object comparison refers to the names of the objects/districts and the localisation of their boundaries : 140 objects on 363 (38%) are correct. Three types of mismatches in-between the datasets are identified (Figure 10). While the Gaul initiative attempts to refer to an international nomenclature and agreement, the naming regulations are more relevant for an international use of the Caucasian database. Table 19 states the number of errors in the three types and the action taken in the consolidation phase of this dataset. For all the Russian part of our dataset, Gaul does no provide any description at the district level, corresponding to an administrative level 3. The administrative level 2 for this dataset is not sufficient for the objectives of our analysis because an homogenous average size of the area of "district" objects is needed. The Caucasian database provides then a higher level of precision than the 11 oblast/ regions in Gaul. In Armenia, the level 2 refer to smaller units within the region, Administrative level 1 correspond to the "district" objects. The combination of these two source of information improve the quality of the district dataset in some part of the region, but in others the district layer is more precise than Gaul.

Table 19 : GAUL and district data comparison

Error	Description	Number	Action Taken
Gaul & District Mismatch	Gaul data shows additional boundaries	21	Gaul description given to underlying polygons
District & Gaul Mismatch	District data shows additional boundaries	191	Description derived from Gaul overlaying polygon
No descriptions	No descriptions in Gaul	11	Gaul Admin 2 descriptions used

Figure 10 : GAUL and district dataset overlay with three types of errors

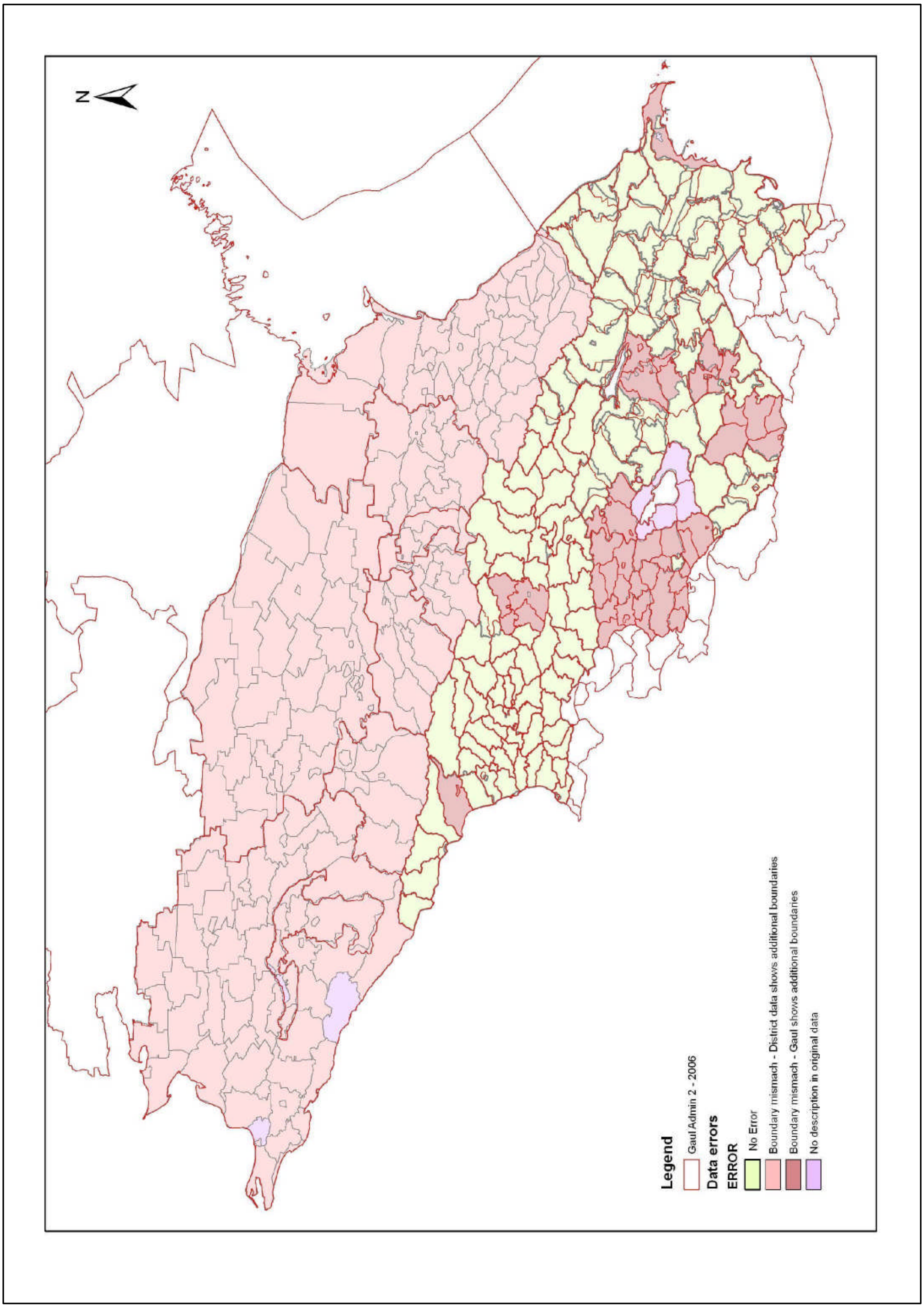

N
Legend
Gaul Admin 2 - 2006
Data errors
ERROR
No Error
Boundary mismach - District data shows additional boundaries
Boundary mismach - Gaul shows additional boundaries
No description in original data

Table 3 states that most of the fields cannot be used. Attribute accuracy checking is limited to identify inconsistencies in the typology of the following layers : forest, rivers, lakes, districts, roads and settlement.

Table 20 summarizes the number of objects per categories for the forest layer. The typology includes blank cells and miscoding that represent 1% of the objects in comparison of the overall area of the forest layer. Unfortunately the definition provided by the producer do not distinguish two types of forest called “fr” and “fs”. For the “river” entity (Table 21), “rr” and “rv” have not been distinguished yet, and “cr” refer to misclassifications . Blank values and “cr” type account for less than 5% attributes errors. The same undefined classes can be seen for the “lake” entity (Table 22), while no typological errors have been noticed in this layer. In the “road” layer Table 23, a lot of miscoding errors, including blank values, have been detected but they do not represent a high percentage of the total length of linear segments (3.3%). The roads digitilised inside the urban or village sprawls (“ul”, “vl”). Table 24 lists “settlement” layer miscoding errors, representing a negligible percentage of coverage (less than 0.01%).

Table 20 : Forest typology and percentage of coverage

TYP	Definition	Cnt_TYP	Forest coverage percentage	Changes
BLANK		83	0.83%	fs
" fs"		1	0.20%	fs
bs	fallow	231	1.44%	
fp	plantation	274	2.24%	
fr	forest	247	1.02%	
fs	forest	3076	94.27%	

Table 21 : River typology and percentage of coverage

TYP	Cnt_TYP	%number	%length
BLANK	22	3.84%	3.59%
cr	44	7.68%	1.22%
rr	2	0.35%	0.14%
rv	505	88.13%	95.04%

Table 22 : Lakes typology and percentage of coverage

TYP	Definition	Cnt_TYP	percentage of area (seas area not included)
gl	glacier	265	35.63%
lk	lake	2265	52.56%
lp	non permanent lake or lagoon	1066	2.08%
rr	river	5	0.19%
rv	river	396	3.53%
rz	artificial lake	607	6.01%
sea	sea	1	
zl	bay	1	

Table 23 : Roads typology and percentage of coverage

TYP		Cnt_TYP	Sum_leng	% of length
el		16	737.0274	0.19%
gdu		1	9.7713	0.00%
np		7	340.7228	0.09%
p[d		1	3.7914	0.00%
pdp		1	9.9987	0.00%
pdpd		1	10.0326	0.00%
pds		2	6.8916	0.00%
pp		34	3247.2572	0.82%
ps		1	3.4347	0.00%
s		1	2.1930	0.00%
tr		640	7919.5174	1.99%
u		3	4.6789	0.00%
uus		1	1.3306	0.00%
uuu		1	0.9115	0.00%
as	motorway	17	300.4895	0.08%
ep	oil pipe	56	4307.2325	1.08%
gg	path	7178	55951.0826	14.06%
gp	gas pipe	114	6163.0053	1.55%
gu	good path	5365	43139.5398	10.84%
pd	forest path	8917	81934.2016	20.59%
ru	road in construction	57	1070.0311	0.27%
rw	railway	1161	31033.5811	7.80%
ss	road	9135	99307.9445	24.96%
su	state road	2801	52447.1317	13.18%
ul	urban road	1004	1700.2513	0.43%
uu	village road	6406	7376.2925	1.85%
BLANK		46	839.4022	0.21%
			397867.7448	
undefined			13136.9613	3.30%
sprawls roads			8215.6947	2.06%

Table 24 : Settlement typology and percentage of coverage

TYP	Definition	Sum_area_c	Percentage
cc	Capital	224.3064	0.80%
ct	Center of region	1635.6572	5.80%
kc	Center of District	578.3205	2.05%
rc	Large town	2376.0838	8.43%
tw	Town	687.9641	2.44%
vl	Villages	22539.8785	79.99%
BLANK	undefined	3.6411	0.01%
Komeremi	undefined	0.2192	0.00%
cl	undefined	1.5256	0.01%
ks	undefined	5.9039	0.02%
os	undefined	30.6236	0.11%
rs	undefined	3.6216	0.01%
rt	undefined	10.7229	0.04%
st	undefined	73.3675	0.26%
te	undefined	5.0924	0.02%
vlvl	undefined	0.1966	0.00%
		28177.1249	

The attribute accuracy element of data quality summarizes the errors in classification related to a true categorisation. The performance of attribute accuracy assessments was limited by the lack of explicit attributes. The ability to assess the attributes available was limited by whether they could be sufficiently verified by an independent source. As the set up of this assessment needs relevant control data, the overall methodology is applied only on the settlement feature in the Nagorno Karabagh area (Figure 11). The percentage of agreement of the Kappa statistic, already used in the positional accuracy assessment, can be applied in this topic. Really bad results (19% agreement) of this small test illustrate the difficulties of the attribute evaluation. Especially in this disputed area, names of settlement are changing and different languages (Armenian and Azeri as well as ethnical dialects) are used. Road assessment on the

Figure 11 : Tertiary data of settlements on Nagorno Karabagh region

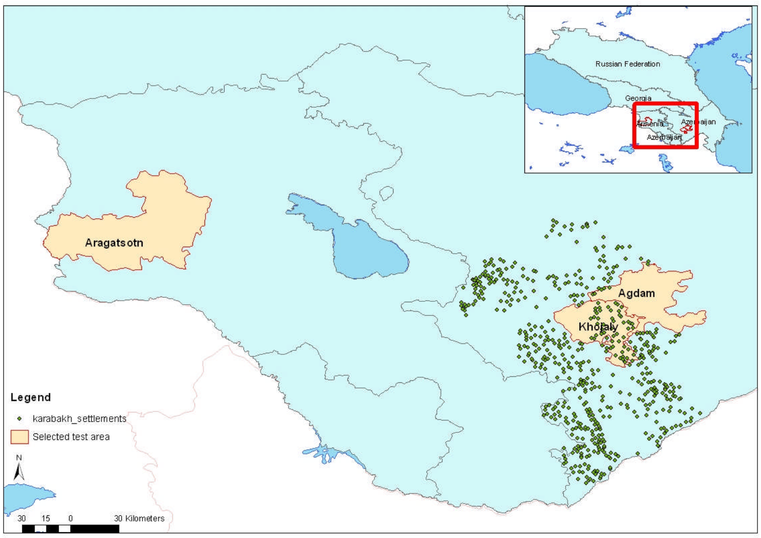


Table 25 : Attribute accuracy test for settlements data

Test Area	Karabakh
Link field	Settlement name
Attribute type	Nominal
Total number of records in test	565
Total number of records matched	106
Percentage of agreement	19%

5.5. Completeness

Like accuracy, completeness can be divided into two components : entity completeness and attribute completeness. Entity or attribute completeness refers to the exhaustiveness of the dataset in terms of the entity type it is intended to represent. Two measures of completeness are needed because of two possible types of errors: omission and commission. Errors of omission occur when a feature in the control data does not have a corresponding feature in the test data. Errors of commission occur when a feature in the test data does not have a corresponding feature in the control data (Table 26). In the particular case, the regional extent and the instability topic create a high level of abstraction from the visual reality. Relevant information to represent the geopolitical instability both for the dataset producer and user cannot be considered as equal to the real features detected on the field or on a very high resolution imagery. In other words, without knowing what the data set is intending to describe, it is difficult to assess completeness.

Table 26 : Feature completeness

Test Data	Control Data	
	Present	Absent
Present	Correct	Error of Commission
Absent	Error of Omission	Correct

Because of low knowledge of the regional geopolitical context, the geomodelling approach supposes that the factors studied in the geopolitical atlas represent the overall picture of the instability driving force. The discussion about the completeness of the particular dataset then refers to the coherence between the supposed list of data and the actual information but also to the spatial coverage of the factors intended to be represented. In a second step, the completeness accuracy refers to the intended model for the user, the geomodel.

5.5.1. Comparison between the list and the existing information.

The statistical dataset acquired by the user from the producer was supposed to contain all the socio-economic themes listed in Table 2 at the dates mentioned. From the invoice proposed, we have to report on missing data for education (number of student) and for employment in all sectors (agriculture, industry and services). Moreover, the geographical coverage of the different themes is not complete, usually because of changes in administrative units or disagreements on conflicting area (Table 27). This table uses the structure of Table 2 and identify the regions where data are missing or questionable. Incomplete information is a major issue in a regional study of instability. In particular when the main disputed areas often present missing values in national statistics sources.

Table 27 : Regional data availability

	Armenia	Azerbaijan	Karabakh	Nkhitshevan	Okkup	Abkhazie	Adjarie	Georgia	South Oss	Adygea	Checnia	Daguestan	Ingushetia	KBR	KCR	Krasnodar	N.Ossetia	Stavropol
Total 1979	🟨	🟨	🟨		🟨		🟨	🟨	🟨						🟨			
Total 1989	🟨	🟨	🟥	🟨	🟥	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨
Total 1995	🟨	🟨	🟨	🟨	🟨		🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨
Total 2002	🟨	🟨	🟥	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨
Total 2005	🟨	🟨	🟨	🟨	🟨		🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨
Urban 1979	🟨	🟨	🟨		🟨		🟨	🟨	🟨						🟨			
Rural 1979	🟨	🟨	🟨		🟨		🟨	🟨	🟨						🟨			
Urban 1989	🟨	🟨	🟥	🟨	🟥	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨
Rural 1989	🟨	🟨	🟥	🟨	🟥	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨
Urban 1995	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨
Rural 1995	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨
Urban 2002	🟨	🟨	🟥	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨
Rural 2002	🟨	🟨	🟥	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨
Eponym 89	🟨	🟨	🟥	🟨	🟥	🟨	🟨	🟨	🟨	🟨	🟥	🟥	🟨	🟨	🟨	🟨	🟥	🟨
Eponym 02	🟨	🟨	🟥	🟨	🟥	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨
Birth 95	🟨	🟨	🟨	🟨			🟨	🟨	🟨	🟨		🟨		🟨	🟨	🟨	🟨	🟨
Death 95	🟨	🟨	🟨	🟨			🟨	🟨	🟨	🟨		🟨		🟨	🟨	🟨	🟨	🟨
Birth 02	🟨	🟨	🟥	🟨	🟥		🟨	🟨	🟨	🟨		🟨	🟨	🟨	🟨	🟨	🟨	🟨
Death 02	🟨	🟨	🟥	🟨	🟥		🟨	🟨	🟨	🟨		🟨	🟨	🟨	🟨	🟨	🟨	🟨
Infant death 95		🟨	🟨	🟨	🟨		🟨	🟨	🟨	🟨		🟨		🟨	🟨	🟨	🟨	🟨
Infant death 02	🟨	🟨	🟥	🟨	🟥		🟥	🟥	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨
Weddings 95		🟨	🟨		🟨		🟨	🟨	🟨	🟨		🟨		🟨	🟨	🟨	🟨	🟨
Divorces 95		🟨	🟨		🟨		🟨	🟨	🟨	🟨		🟨		🟨	🟨	🟨	🟨	🟨

Weddings 02	🟨	🟨	🟪	🟨	🟪	⬜	🟨	🟨	🟨	🟨	⬜	🟨	⬜	🟨	🟨	🟨	🟨	🟨
Divorces 02	🟨	🟨	🟪	🟨	🟪	⬜	🟨	🟨	🟨	🟨	⬜	🟨	⬜	🟨	🟨	🟨	🟨	🟨
Area cadastr 02	🟨	🟨	🟪	🟨	🟪	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨
Agric. areas 02	🟧	🟧	🟪	🟧	🟧	🟧	🟨	🟨	🟨	🟨	🟧	🟨	🟨	🟨	🟧	🟧	🟨	🟧
All sown areas02	🟨	🟨	🟪	🟧	🟧	🟧	🟨	🟨	🟨	🟨	🟧	🟨	🟨	🟨	🟨	🟨	🟨	🟨
Plantations02	🟨	🟨	🟪	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟧	🟨	🟧
All sown areas 95	🟨	🟨	🟨	🟨	🟨	⬜	⬜	🟨	🟨	🟨	⬜	🟨	⬜	🟨	🟨	🟨	🟨	🟨
Plantations 95	🟨	🟨	🟨	🟨	🟨	⬜	⬜	🟨	🟨	🟨	⬜	🟨	⬜	🟨	🟨	🟨	🟨	🟨
Pastures 02	🟨	🟨	🟪	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨
Forests 02	🟨	🟨	🟪	🟨	🟨	🟧	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟧	🟨	🟧
Pastures 95	🟨	🟨	🟨	🟨	🟨	⬜	⬜	🟨	🟨	🟨	⬜	🟨	⬜	🟨	🟨	🟨	🟨	🟨
Forests 95	🟨	🟨	🟨	🟨	🟨	⬜	⬜	🟨	🟨	🟨	⬜	🟨	⬜	🟨	🟨	🟨	🟨	🟨
Cereals areas02	🟨	🟨	🟪	🟨	🟨	⬜	🟨	🟨	🟨	🟨	⬜	🟨	⬜	🟨	🟨	🟨	⬜	🟨
Cereals yields 02	🟨	🟨	🟪	⬜	🟨	⬜	🟨	🟨	🟨	🟨	⬜	🟨	⬜	🟨	🟨	🟨	🟨	🟨
Cereals areas95	🟨	🟨	🟨	🟨	🟨	⬜	🟨	🟨	🟨	🟨	⬜	🟨	⬜	🟨	🟨	🟨	🟨	🟨
Cereal yields 95	🟨	🟨	🟨	🟨	🟨	⬜	🟨	🟨	🟨	⬜	⬜	🟨	⬜	🟨	🟨	🟨	🟨	🟨
Cattles 95	🟨	🟨	🟨	⬜	🟨	⬜	🟨	🟨	🟨	🟨	⬜	🟨	⬜	🟨	🟨	🟨	🟨	🟨
Cattles02	🟨	🟨	🟪	🟨	🟨	⬜	🟨	🟨	🟨	🟨	⬜	🟨	🟨	🟨	🟨	🟨	🟨	🟨
Doctors02	🟨	🟨	🟪	🟨	🟨	⬜	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨
Doctors 95	🟨	🟨	🟨	🟨	⬜	⬜	🟨	🟨	🟨	🟨	⬜	🟨	⬜	🟨	🟨	🟨	🟨	🟨
Phones 02	⬜	🟨	🟪	🟨	⬜	⬜	⬜	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨
Cars02	🟨	🟨	🟪	⬜	⬜	⬜	🟨	🟨	🟨	⬜	🟨	🟨	🟨	🟨	🟨	🟨	🟨	⬜
Cars 95	🟨	🟨	🟨	🟨	⬜	⬜	🟨	🟨	🟨	🟨	⬜	🟨	⬜	🟨	🟨	🟨	🟨	🟨
Monthly Wage02	🟨	🟨	🟪	⬜	🟨	⬜	⬜	⬜	⬜	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨
Monthly Pension02	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨
Industry product02	🟨	🟨	🟪	🟨	🟨	⬜	🟨	🟨	🟨	🟨	⬜	🟧	⬜	🟨	🟨	🟨	🟨	🟨
Agric. product 02	🟨	🟨	🟪	🟨	🟨	⬜	⬜	⬜	⬜	🟨	⬜	🟨	⬜	🟨	🟨	🟨	🟨	🟨

⬜	no data
🟨	available data
🟪	unreliable
🟧	restored

The comprehensive geopolitical picture provided by the atlas project refer to an extensive list of themes (Table 8). Some of these 28 factors are not available at the district level but at national or regional scale (39%) (Table 28). The spatial disaggregation of statistics at a higher resolution is one of the techniques that will be tested in the geomodeling approach (See chapter 7).

Table 28 : Available information from instability atlas

Topics	Instability factors	In progress	Other Scale
History	Different definition of Caucasian territories	X	X
	"Dreamed" or historical territories	X	X
	Evolution of the administrative units	X	X
History of conflicts	Caucasian conflicts since 1988	X	X
	Territorial and border contests		
Population and Demography	Population density by district (hab/km2)		
	Urban Population (%) and city size		
	Population growth (189/2002/2005)		
	Birth ; Death ; Natural growth by district		
	Infant mortality		
	Migration, regional	X	X
The ethnic mosaic	Ethno-linguistic Map of the Caucasus		
	Religions		
	First Nationality -nation 1-, by district 2002		
	Second nationality -nation 2-, by district 2002		
	Eponym population by district 1989/ 2002		
	Russians population evolution		
Economy	PIB per capita and evolution in time		X
	Active population and unemployment per region		X
	Main economic projects	X	
	Electricity (production and consumption)		X
	Industry per capita		
	Main industrial plants	X	
	Agricultural production per capita, by district		
	Wood and wood processing		
Social development and disparities	Poverty		X
	Population equipment (phones and cars)		
	Salary / pension		
	Health sector equipment		
	Crime		
Transport and foreign trade	TRACECA; Tubes and oil transport		
	The ways of traffic (air and airports)		X
	Foreign trade by States		X

The district spatial dataset provided do not cover some topics (25%), as for example the “dreamed territories or history of conflict”. This information is currently made available, often at another scale, in the static format of maps. The development of the updated version of the 1997’ Atlas map is not completed yet. These highly relevant information in the instability perspective will be integrated in the geospatial data model with the collaboration of the producers. This information will then be integrated in the JRC comprehensive geodatabase.

5.5.2. Comparison with the scope of the geomodel

The geomodel of instability refers to quantitative conflict analysis addressing the territory as one of the most important explanatory variable. In Political Geography or International Relations scholarships, territorial dimensions includes contiguity or proximity, nature of borders, distribution and diffusion in space and time of socio-economical processes and resources uses (Table 29). This table list categorical indicators used at country level to conduct a global analysis.

In literature, there is a growing consensus that to be able to study current conflict – that are, by and large, civil conflicts (O'Loughlin 2005, Restrepo et al. 2005) – variable should be available at sub-national level (Hauge and Ellingsten 1998). The study of civil war using country-level statistics is deemed to be "potentially flawed" (Buhaug and Rod, 2005) because country level statistics "dilute" the importance of determinants of violence that occur at local level. Based on this statement, the geomodel has to apply and adapat this insecurity frame in the security complex region of the Caucasus.

Table 29 : Literature on conflict reviewed and sorted by spatial concepts

Spatial concept	Indicator name	References
Diffusion of conflict	previous conflicts	Urdal 2005
	regime type / level of democracy	Urdal 2005, Collier and Hoeffler 1998, Fearon and Laitin 2003
Proximity	relevant neighbours or contiguous states	Richardson 1960, Diehl 1999
	distance between capitals	Lemke 1995
	distance between centroids	Richardson 1960, Vanzo 1999
	minimum distance threshold	Gleditsch and Ward 2001
Border effect	the number of shared borders	Wesley 1962
	length of borders	Wesley 1962, Furlong et al. 2006
	type of border – homeland or colonies-	Starr and Thomas 2002
	cost in time necessary to cross the border according to topographical elements on the border	Bueno de Mesquita and Lalman 1992
	technological changes of this cost in time	Lemke 1995
	salience or willingness	Senese 1999
	ease of interaction or opportunity using a GIS approach (road, railway, terrain steepness, population, infrastructures within a 4 km buffer)	Starr 2002
Population	total number	Wils et al. 1998
	population density	Hauge and Ellingsten 1998
	population growth	Urdal 2005
	percentage of inhabited region related to the land cover	Buhaug and Rod, 2005
Inequality	fractionalization (ethnic, economic and social)	de Soysa 2002, Fearon and Laitin 2003
	urban population	de Soysa 2002, Homer-Dixon 1999
	fragmentation/polarization (ethno-socio-economic)	Buhaug and Gates 2002
	poverty	Collier and Hoeffler 1998
	infant mortality	Sen 1998
	income inequality (GINI)	Collier and Hoeffler 1998, Fearon and Laitin 2003, Murshed and Gates 2005
External Influences	international trade	Gleditsch 2002, de Soysa 2002
Environment	cropland	Wils et al. 1998, Urdal 2005
	land degradation	Hauge and Ellingsten 1998
	roughness of the terrain	Bueno de Mesquita and Lalman 1992, Lemke 1995, Starr 2002, Fearon and Laitin 2003 / but non relevant for Collier and Hoeffler 1998, Buhaug and Gates 2002
Natural resource availability	primary commodities	Collier and Hoeffler 2004, de Soysa 2002, Elbadawi and Sambanis 2002
	presence of resources including oil, gemstones, illicit crops	Fearon and Laitin 2003, Ross 2004
	diamonds	Lujala et al. 2005
	timber	Ross 2006
	freshwater availability	Hauge and Ellingsten 1998, Toset et al. 2000, Furlong et al. 2006

The Caucasus study will explore two distinct modelling approaches: (i), building a spatial and continuous muticriteria model of instability integrating in a continuous GIS the geopolitical factors (ii) defining subnational values (at the district level) of the list of indicators for the Caucasus region (See chapter 7). For these both modelling perspectives, the complete list of instability factors will refer to this literature review, as well as the geopolitical picture provide by the Atlas. As the final quality of the geomodel is linked to availability and reliability of data, this report was a necessary initial step before the model set up.

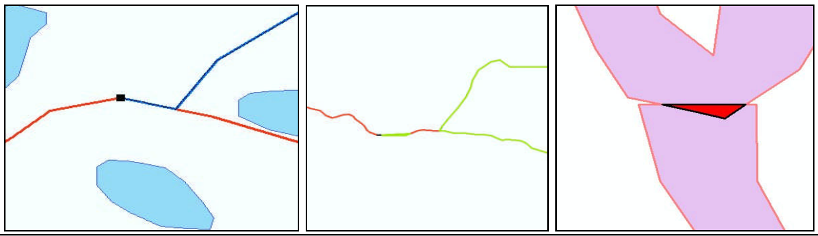
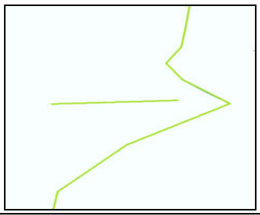
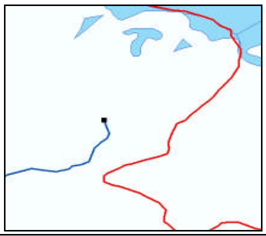
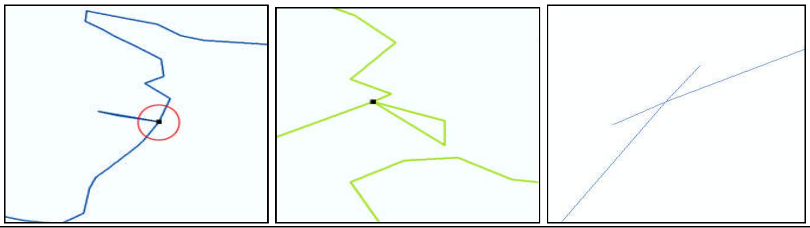
5.6. Consistency

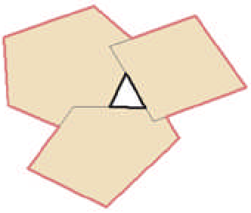
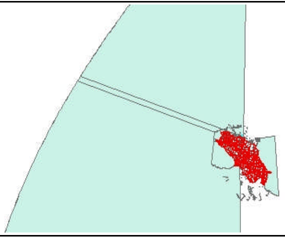
Consistency as a general term deals with logical rules of the structure and relationships between data in the database.

5.6.1. Geometrical consistency

Arc/Info GIS software was used to check the logical consistency of the topology of the data sets. Data was checked for a list of topological errors (Table 30) : (i) duplicated (self overlapping) lines representing the same entity (ii) linear segments or polygons not appearing to be part of an object or not put there intentionally (usually by using a snap tolerance), (iii) dangles created when digitized linear objects stop short of, or extend past, an intended intersection point, (iv) intersecting lines or pseudo nodes that did not represent island polygons (v) gaps in between polygons, (vi) sliver polygons sometimes created when duplicated lines have not been removed.

Table 30 : Topological errors checked in ArcGIS

Errors	Description
Self-Overlapping	Duplication of lines or polygons 
Single Part features	Small fragment lines or polygons (tolerance) 
Dangles	End point not connected to the line 
Self-Intersecting or Pseudo-nodes	Lines looping back forming small error polygons 

<table>
<tr><td>Gaps</td><td>Voids within a polygons or between adjacent ones
</td></tr>
<tr><td>Sliver polygons</td><td></td></tr>
</table>

Topological rules are used to estimate these errors in the Caucasian dataset (Table 31):
- "Must Not Overlap" requires that *line and polygons features* not overlap themselves. Lines can cross or touch themselves, but must not have coincident segments. The *polygons* can share edges or vertices but the interior of polygons in the feature class not overlap. This rule is used when an area cannot belong to two or more polygons.
- "Must Not Have Gaps" requires that there are no voids within a single *polygon* or between adjacent polygons. All polygons must form a continuous surface. An error will always exist on the perimeter of the surface. You can either ignore this error or mark it as an exception. Use this rule on data that must completely cover an area. For example, soil polygons cannot include gaps or form voids—they must cover an entire area.
- "Must not have dangles" requires that a line feature must touch lines from the same feature class at both endpoints. An endpoint that is not connected to another line is called a dangle. This rule is used when line features must form closed loops, such as when they are defining the boundaries of polygon features. It may also be used in cases where lines typically connect to other lines, as with streets. In this case, exceptions can be used where the rule is occasionally violated, as with cul-de-sac or dead end street segments.

Table 31 : Summary of topology errors in dataset

Data	Feature Type	Must not intersect	Must not Overlap	Must not have gaps	Total
Ethnic	Polygon		424	1161	1585
District	Polygon		924	1386	2310
Settlements	Polygon		109		109
Lakes	Polygon		301		301
Forest	Polygon		408		
Rivers	Line	0	8		150
Railway	Line	949	784		1733
Pipes	Line	81			81
Main Roads	Line	6175	2204		8379

5.6.2. Attribute consistency

As explained in the description of the dataset (See 5.1.4) two types of attributes are distinguished and consequently two types of attribute consistency check can be envisaged. Used in this report as an example of the quantitative quality evaluation, the settlement attributes consistency is evaluated in its hierarchical structure. The settlement feature attribute typology will be checked by analyzing the 6 categories of urban sprawls based on rules as their belongings to the region or district from which they are the center, comparison with external population sources and sprawl area.

Table 32 : settlement categories

Capital (Baku, Yerevan + 2 Tbilissi)	1
Head of Region in Armenia, Georgia, Nakhichvan,	2
Head of District	3
Town with more than 50000inhab	4
Town	5
Villages	6

For each district, the internal inconsistencies of the statistical layer (socio-economic data at district level) are identified through the following list of rules :

- Statistical area of statistical unit = GIS area of statistical unit (with a tolerance level)
- Population >< 0 and >0
- Total population= Rural + Urban populations
- Total of ethnic groups = total population

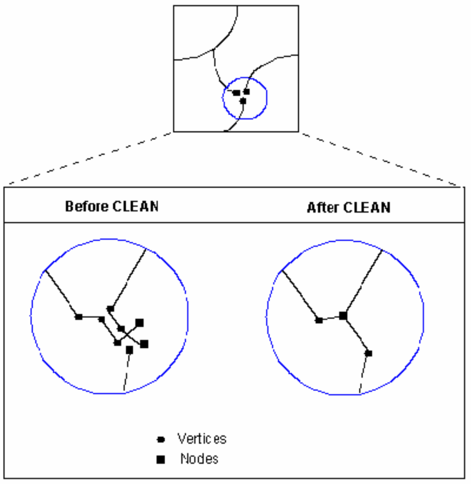
To check the consistency of the district data, they can also be aggregated to a lower administrative level to be compared to other data sources (further check and data comparison).

6. Data cleansing and improvement

6.1. Topology cleansing

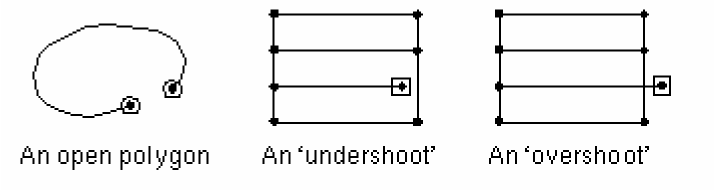
ArcInfo Workstation was used for cleansing the data as it provided robust and automated functionality for identifying and resolving errors. ArcInfo workstation provided 2 main functions for resolving the topology errors. The “CLEAN ” tool which generates a coverage with correct polygon or line topology (Figure 12). To do this, CLEAN edits and corrects geometric coordinate errors, assembles arcs into polygons and creates feature attribute information for each polygon or arc. Clean allows the user to specify the distance within which a new arc will be extended to intersect an existing arc also known as the snapping tolerance. CLEAN also eliminates all duplicate or overlapping lines or polygons.

Figure 12 : Clean tool in ArcInfo workstation



Once CLEAN has been run on the data remaining errors can be identified graphically using ArcEdit (a component of ArcInfo workstation). In ArcEdit dangles in the data can de drawn; dangles help identify errors in the data like overshoots, undershoot and open polygons. These errors have to be resolved manually using tools provided in Arcedit.

Figure 13 : Dangles errors identified by ArcEdit tool in ArcInfo workstation



The cleaned datasets were then imported into a Geodatabase and checked against topology rules, this to ensure that the topology achieved was clean and thorough. Table 33 introduces these rules details their use. The amount of time necessary to carry out this cleansing phase help to quantify the error content of the dataset (Table 34)

Table 33 : Topological errors cleansing processes

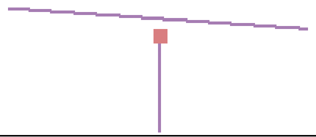
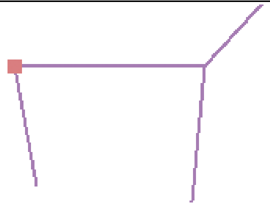
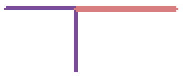
Topology rules	Potential fixes	Comments
Must Be Larger Than Cluster Tolerance (POLY)	Delete	Any polygon feature would collapse when the cluster tolerance level is reached
Must Not Overlap (POLY)	Subtract, Merge, Create Feature	
Must Not Have Gaps (POLY)	Create Feature	Create a new polygon in the void in between or mark the error on the outside boundary as an exception.
Must Not Have Dangles (LINES)	Extend, Trim, Snap	
Must Not Have Pseudonodes (LINES)	Merge to Largest, Merge	
Must Not Self Overlap (LINES)	Simplify	

Table 34 : Data Cleansing processes (ArcInfo workstation) and time assessment

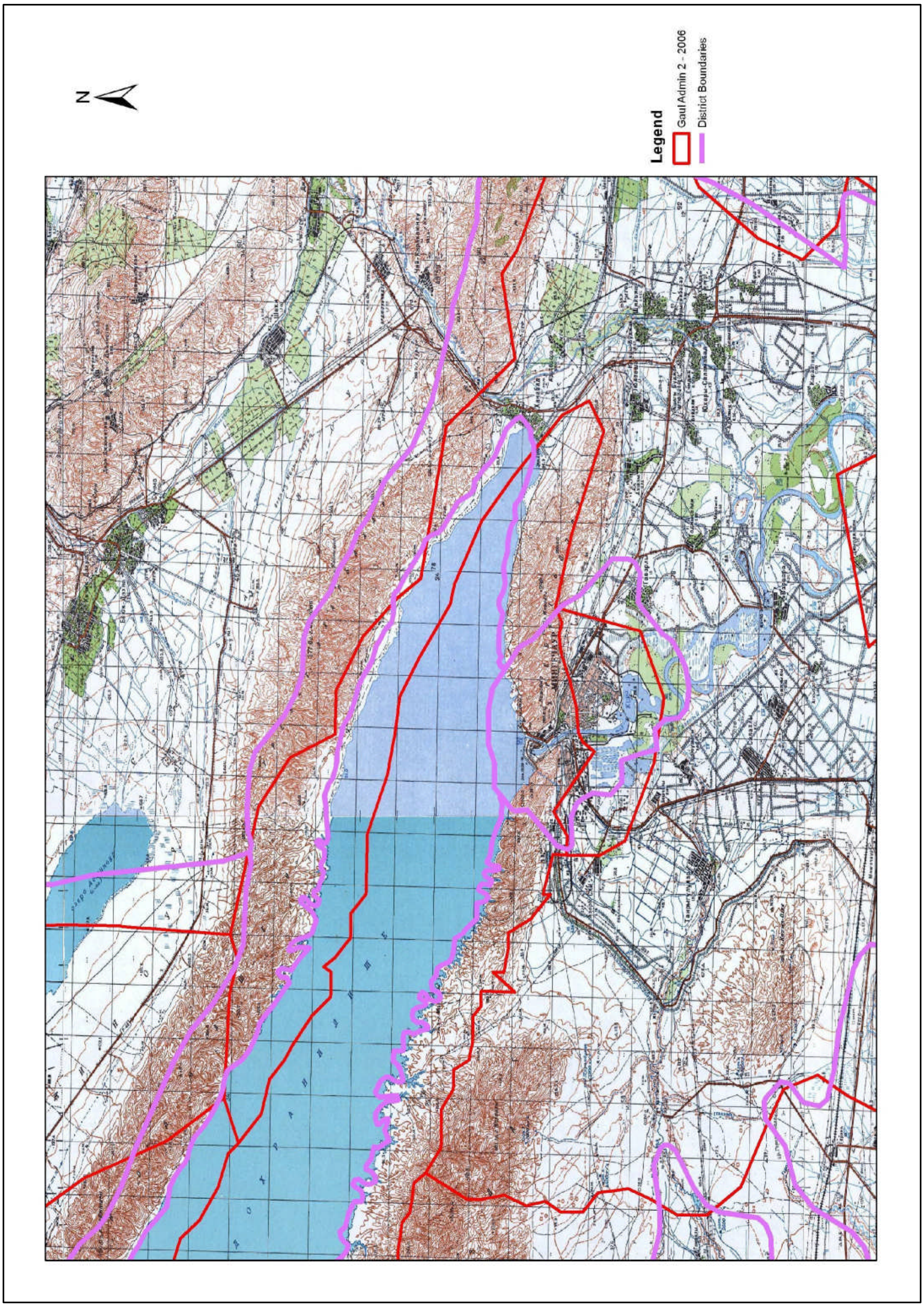
Data	Clean/build	Resolve dangles	Check and resolve topology problems	Snap Tolerance (decimal degrees)	Completion	Total time
Ethnic	2	8	5	0.01	all	15
District	3	6	7	0.0001	all	16
Settlements	1	3	2	0.0001	all	6
Lakes	1	4	2	0.0001	all	7
Forest	3	4	2	0.0001	all	9
Rivers	1		1	0.0001	Not all dangles resolved as river flow & network are unknown	2
Railway	1		2	0.0001	Not all dangles resolved, begin and end points of rails unknown	3
Pipes	1	1	2	0.0001		4
Main Roads	2		2	0.0001	Not all dangles resolved begin and end points of roads unknown	4
					Total	66

6.2. Data consolidation and improvement

As described several time in the corpus of this report, different methods have assessed the errors and inconsistencies of topographical entities and associated attribute tables. The main improvement are related to (i) the topological cleansing process explained before, (ii) the district and Gaul comparison (topology and attribute), and (iii) the improvement of the matching between ethnic and district dataset

For district data, a comparison with Gaul data set was used to update and improve the attribute table. Table 19 shows the problems experienced but also how they were resolved. A number of areas showed large geographical differences when compared. These areas were overplayed over a 100k Soviet topographical map and showed that these data shift were more evident in the Gaul boundaries (Figure 14). Boundaries with best comparison to the Soviet maps were included in the final dataset.

Figure 14 : Visual comparison between Gaul and district data



The cleaned Ethnic data had an outer boundary was inconsistent and over-generalised. In order to have uniformity between the datasets it was decided to use the outer boundary of the final districts data and replace this with the existing Ethnic outer boundary. This was done by dissolving all the boundaries in the district data. The resulting dataset was a single outline of the districts data. The existing outer boundary was erased and ArcInfo/Edit was used combine the 2 features and resolve any topology errors.

6.3. Attribute cleansing

The cleansing process also identified attribute errors. The following were identified as common attribute errors within the data:

- Missing descriptive information
- Misspelling
- Duplicate Record

Table 35 shows examples of the attribute errors evident in the district data and how they were resolved. For records that were blank or were incorrectly duplicated, reference was made to the relevant control data. The main attribute table refers to the district dataset, the application of consistency rules results in a new geodatabase associated with a large statistical dataset that can be used for different purposes (producer and user).

Table 35 : Examples of attribute errors

Attribute problem	Problem resolved	Description	Solution
Armavir		2 or more incorrect shapes included in a single boundary. Sliver polygons incorrectly adopting district name.	Merge Records
Armavir	Armavir		
Armavir			
		Record entry contains incorrect spelling, numeric symbols or has different case settings	Correct incorrect symbols, spelling and case. control datasets used as reference
armavir_	Armavir		

7. Future spatial processing & analysis

Based on the result of this report, the Caucasus modelling study will explore two distinct modelling approaches already stated in the completeness analysis:

- a spatial and continuous muticriteria model of instability integrating in a continuous GIS the geopolitical factors
- defining instability indicators values for subnational spatial entities (district units) throughout the Caucasus region.

7.1. SDSS model

The first multicriteria approach will use a GIS continuous mapping approach to standardize the criteria in a SDSS. The homogenous GIS framework means that some specific choices have to be discussed : (i) an unambiguous spatial reference system (Afgoye, UTM38N), (ii) a continuous raster grid (grid cells density rather than lines), and (iii) an uniform range of values addressing the “suitability” in terms of “instability” for the societal security.

Challenging steps in this process refer to the building of continuous datasets while the statistical data are available at the district level. The most interesting procedure to disaggregate spatially these data could use a “population density mapping method”. The settlement dataset could be used to disaggregate the population census data at a pixel resolution of 1km. Linking the socio-economic information available at the district level to this demographic raster layer is an attempt to “socialise” the pixel.

Spatial analysis methods based on various forms of distance (Euclidean, cost, time) which are generally quantitative and continuous will also be considered. The potential accessibility (Wegener et al. 2002) (isotropy, homogeneity) can be the basic map to assess the effectiveness of borders or roads/railway networks. The space can also be seen as polarized by nodes (polycentric urban and transportation network). The permeability maps (Stephenne and Pesaresi, 2006) can either be used as input, as well as an illustration of the SDSS approach.

The functionality of the GIS can be extended to facilitate complex analysis of spatial features as in some application of GIS analysis like (i) hydrological planning, (ii) transport planning, or (iii) urban or land use planning. GIS has become a particularly useful and important tool in hydrology and to hydrologists in the scientific study and management of water resources. Because water in its occurrence varies spatially and temporally throughout the hydrologic cycle, its study using GIS is especially practical. Network and land data can be easily developed, maintained and updated in GIS database. Transportation planning and management needs accurate and timely spatial and non-spatial information like, network, capacity, speed restriction etc., to assist planning activities. Most importantly GIS database maintains and provides topological relationship (connectivity and contiguity), which plays key role either in macro or micro level transportation planning analysis. Standardisation and Data Sharing are the two components provided strong support for implementing the enterprise GIS in urban transportation planning. The analysis capabilities of a GIS package allow the urban planner to address what-if questions and work out a variety of action plans in a scientific manner. A number of problems can be solved by geographic analysis (town ship development , relationships between agricultural parameters such as yield and salinity, land capability analysis, site locations for facilities, environmental problems such as animal migration.

Spatial processing techniques can integrate continuous datasets derived from remote sensing source with the Caucasus geodatabase, as for example,
- flow of the rivers (combining the SRTM with the river dataset)
- potential landslide (combining the SRTM - average steep - with the forest coverage and the potential impact on the infrastructure / settlement / agriculture)
- pollution sources : combine the pollution point file (ENVISEC information associated to the settlement layer using topographic maps as control data) with the river information to define the flows of river that are potentially polluted.

7.2. District level instability

The second approach will use the district units as the spatial reference. Referring to the list of insecurity indicators discussed in the conflict literature, the geodatabase statistics will be transformed in explanatory criteria for this territorial units. For these both modelling perspectives, the complete list of instability factors will refer to our literature review, as well as to the geopolitical picture provided by the Atlas. These values can then be visually and statistically analysed through model simulations and alternative scenarios using exploratory visualisation facilities. Issues in the spatial disaggregation of national values have to be thoroughly analysed in this future work.

8. Ackowledgement

We would like to thank Igor Savin, Mars-Food/Agri-Fish unit/IPSC for his useful comments on this report. The work was made possible through the support of staff in the Support to External Security unit at the JRC.

9. Discussion and conclusion : overall quality assessment

The bulk of this report has aimed to illustrate how spatial data from various sources have been collected and made ready for use within a GIS. The different evaluation tests allow to give an overall estimation of the dataset quality. This type of data cannot be used at a scale higher than approximately 1:500 000. This Caucasian dataset has the objective to provide an overall picture of the regional security complex and not a precise localisation of specific real features. This fact has to be kept in mind in the following processing modelling stages.

This report also provides an adapted methodology to assess quantitatively the quality of a database with no metadata information. The elements of data quality are envisaged in a progressive way in this report and thoroughly studied for the settlement layer. The other layers are evaluated in a less in-dept way but allow the test of different methods associated to the three types of features (point, line, polygon).

The ongoing management of GIS data should include methodology for the improvement of the data accuracy. As newer datasets like satellite imagery can be integrated in the database, the spatial processes must be done in a way to ensure the new data has compatibility with the defined geo model. As finer scales are introduced, the accuracy can be increased so is the capability of the GIS as a whole improved. These processes may include:

- Digitization from raster data
- Geoprocessing like Clip, buffer, update , aggregate etc
- Clean & building of data
- Topology checks and corrections

10. References

Art, R., 1993, Security, In *The Oxford Companion to Politics of the World*, J. Krieger (Eds.), pp. 820-822 (New York – Oxford: Oxford University Press, 1993).

Brauch, H. G., 2005, Environment and Human Security. Towards Freedom from Hazards Impacts., InterSecTions publication Serie, 2/2005 (Bonn: United Nations University (UNU-EHS)). http://www.ehs.unu.edu/file.php?id=64.

Bueno de Mesquita, B. and Lalman, D., 1992, *War and Reason,* 322 p. (New Haven: Yale University Press).

Buhaug, H. and Gates, S., 2002, The Geography of Civil Wars, *Journal of Peace Research,* **39**, 4 pp.417-433.

Buhaug, H. and Rød, J. K., 2005, Local Determinants of African Civil Wars, 1970-2001. In *46th annual ISA Convention*, 1-5 March, Honolulu, HI: Norwegian University of Science and Technoloy (NTNU) (www.eldis.org)).

Collier, P. and Hoeffler, A., 1998, On Economic Causes of Civil War, *Oxford Economic Papers,* **50**, 4 pp.563-573.

Collier, P. and Hoeffler, A., 2004, Greed and Grievance in Civil War, *Oxford Economic Papers,* **56**, pp.563-595.

Davies, J., 2005, Uncle Joe Knew Where You Lived (Part I), *Sheetlines from Charles Close Society,* **72**, April.

de Soysa, I., 2002, Ecoviolence: Shrinking Pie or Honey Pot?, *Global Environmental Politics,* **2**, 4 pp.1-36.

Elbadawi, I. and Sambanis, N., 2002, How Much War Will We See?: Explaining the Prevalence of Civil War, *Journal of Conflict Resolution,* **46**, pp.307 - 334.

Fearon, J. D. and Laitin, D. D., 2003, Ethnicity, Insurgency, and Civil War., *American Political Science Review,* **97**, 1 pp.75-90.

Furlong, K., et al., 2006, Geographic Opportunity and Neomalthusian Willingness: Boundaries, Shared Rivers, and Conflict, *International Interactions,* **32**, 1 pp.79 - 108.

Gleditsch, K. S., 2002, *All International Politics Is Local: The Diffusion of Conflict, Integration, and Democratization,* 266 p. (An Arbor: The University of Michigan Press).

Gleditsch, K. S. and Ward, M., 2001, Measuring Space: A Minimum-Distance Database and Applications to International Studies, *Journal of Peace Research,* **38**, 6 pp.739-758.

Hauge, W. and Ellingsen, T., 1998, Beyond Environmental Scarcity: Causal Pathways to Conflict, *Journal of Peace Research,* **35**, 3 pp.299-317.

Homer-Dixon, T., 1999, *Environment, Scarcity, and Violence,* 253 p. (Princeton, NJ, USA: University Press).

Johnston, D. M., et al., 1999, Quality Assurance/ Quality Control Procedures for Itam Gis Databases, (Geographic Modeling Systems Lab, University of Illinois at Urbana/Champaign). http://www.gis.uiuc.edu/research/spatialanalysis/quality%20assurance.htm, (Last access on 2007 4/10)

Lemke, D., 1995, The Tyranny of Distance: Redefining Relevant Dyads., *International Interactions,* **17**, pp.113-126.

Longley, P. A., et al. (Eds.), 1999, *Geographic Information Systems and Science,* p. (New York: John Wiley & Sons, Ltd.).

Lujala, P., et al. , 2005, A Diamond Curse?: Civil War and a Lootable Resource, *Journal of Conflict Resolution,* **49**, pp.538 - 562.

Murshed, M. and Gates, S., 2005, Spatial-Horizontal Inequality and the Maoist Insurgency in Nepal, *Review of Development Economics,* **9**, 1 pp.121-134.

Richardson, L. F., 1960, *Statistics of Deadly Quarrels,* 375 p. (Pittsburg/Chicago: The boxwood Press / Quadrangle books).

Ross, M. L., 2004, How Does Natural Resource Wealth Influence Civil War? Evidence from 13 Cases, *International Organization*, **58**, pp.35-67.

ROSS, M. L., 2006, A Closer Look at Oil, Diamonds, and Civil War, *Annual Review of Political Science,* **9**, pp.265-300.

SEN, A., 1999, *Development as Freedom,* 366 p. (New York: Alfred A. Knopf).

SENESE, P. D., 1999, Geographical Proximity and Issue Salience : Their Effetcs on the Escalation of Militarized Interstate Conflict, In *A Road Map to War. Territorial Dimensions of International Conflict*, P. F. Diehl (Eds.), pp. 147-178 (Nashville and London: Vanderbilt University Press, 1999).

STARR, H., 2002, Opportunity, Willingness and Geographic Information Systems (Gis): Reconceptualizing Borders in International Relations, *Political Geography,* **21**, 2 pp.243-261.

STARR, H. and THOMAS, D. G., 2002, The 'Nature' of Contiguous Borders: Ease of Interaction, Salience, and the Analysis of Crisis, *International Interactions,* **28**, pp.213-235.

STEPHENNE, N. and EHRLICH, D., submitted, Spatial Concepts and Geo-Technologies in Territorial and Border Analysis: A Review. In *Workshop GMES: Global Monitoring for Environment and Security,* 7 to 8 June 2007, Bolzano, Italy (Caetano: Earsel eproceedings).

STEPHENNE, N. and PESARESI, M., 2006, Spatial Permeability Model at the European Union Land Border, EUR report, 22332 (Luxembourg: European Commission / DG-JRC / IPSC).

TOSET, H. P. W., et al., 2000, Shared Rivers and Interstate Conflict, *Political Geography,* **19**, 8 pp.971-996.

URDAL, H., 2005, People Vs. Malthus: Population Pressure, Environmental Degradation, and Armed Conflict Revisited, *Journal of Peace Research,* **42**, 4 pp.417-434.

VANZO, J. P. F., 1999, Border Configuration and Conflict : Geographical Compactness as a Territorial Ambition of States, In *A Road Map to War. Territorial Dimensions of International Conflict*, P. F. Diehl (Eds.), pp. 73-112 (Nashville and London: Vanderbilt University Press, 1999).

WEBB, P. and HARINARAYAN, A., 1999, A Measure of Uncertainty: The Nature of Vulnerability and Its Relationship to Malnutrition, *Disasters,* **23**, 4 pp.292-305.

WESLEY, J. P., 1962, Frequency of Wars and Geographical Opportunity, *Journal of Conflict Resolution,* **6**, 4 pp.387-389.

11. Terms used

Object: A digital representation of a particular instance of an entity, e.g., a road vector, or stream, or a point rural drop location.

Data Set: A digital collection of objects, e.g., the set of roads, the set of streams. A data set typically represents entities having the same structure and description.

Data Base: A digital collection of Data Sets, with associated methods for querying data sets.

Test Data: A set of objects drawn from a data set to be used to estimate the quality of the data set.

Entity: A real world phenomena, e.g., road, stream, rural drop location.

Control Data: A set of objects drawn from a data set or collected in the field that serves as the standard of comparison for the test data.

12. List of tables

13. List of Figures

European Commission

EUR 22949 EN – Joint Research Centre – Institute for the Protection and Security of the Citizen
Title: Building a spatial decision support system for conflict prevention in Caucasus. Harmonization of heterogeneous sources and data quality assessment procedures.
Author(s): Stephenne Nathalie, Chris MacDonald
Luxembourg: Office for Official Publications of the European Communities
2007 – 51 pp. – 21 x 29.7 cm
EUR – Scientific and Technical Research series – ISSN 1018-5593
ISBN 978-92-79-07140-9

Abstract

Geodata analysis at regional level integrates inevitably some datasets from various sources (statistical, geographical, environmental,…), various scale (regional, national, ..) and various quality: While political structures are constantly changing, as in a potentially conflicting region such as Caucasus, these data integration issues increase. Implementation of quality control methods is an initial and essential step in the integration of geodata inside a spatial regional model. This report provides tools for data harmonization that can be applied to other datasets and other region when existing data sources do not evaluate the quality of their information.

The goal of this report is to provide a quality assessment of the Caucasian GIS dataset to build the Caucasus geomodel of instability/stability. This report evaluates qualitatively and quantitatively the adequacy of this dataset to the objective in following a structured quality assessment protocol (Johnston et al. 1999) and consolidates a final geodatabase. Integrating data from a multitude of derivative geospatial products produced by different sources pose severe problems. Challenges are also introduced by the GIS technology itself. Various data are introduced in this study but the main source of statistical and spatial information is the acquisition of the geopolitical atlas dataset, the “Caucasian dataset” (Radvanyi, INALCO, 2006).

In this report, four data quality elements are identified and described in the specific case of the Caucasian dataset. Lineage information, the three accuracy dimensions (positional, temporal and attribute), logical consistency and completeness evaluations are qualitatively and quantitatively assessed by various metrics. This paper illustrates the use of automatic cartographic and data cleanup techniques of Geographic Information System (GIS) to solve data issues (self overlapping, dangles, pseudonodes and gap in spatial data). This report can further be used as a reference for both the producer and the user to somewhat replace the missing metadata information. Clear statements on dataset quality allow to better communicate in a common goal of understanding the geopolitical Caucasus context.

The bulk of this report has aimed to illustrate how spatial data from various sources have been collected and made ready for use within a GIS. The different evaluation tests allow to give an overall estimation of the dataset quality. This type of data cannot be used at a scale higher than approximately 1:500 000. This Caucasian dataset has the objective to provide an overall picture of the regional security complex and not a precise localisation of specific real features. This fact has to be kept in mind in the following processing modelling stages.

Based on the results of this report, especially the completeness and fitness of the dataset to represent the scope of the model, the Caucasus study will further explore two distinct modelling approaches: (i) a spatial and continuous muticriteria model of instability integrating in a continuous GIS the geopolitical factors, (ii) defining instability indicators values for subnational spatial entities (district units) throughout the Caucasus region.

This report provides an adapted methodology to assess quantitatively the quality of a database when no metadata information is available. The elements of data quality are envisaged in a progressive way in this report and thoroughly studied for the settlement layer. The other layers are evaluated in a less in-depth way but allow the test of different methods associated to the three types of features (point, line, polygon).

The mission of the JRC is to provide customer-driven scientific and technical support for the conception, development, implementation and monitoring of EU policies. As a service of the European Commission, the JRC functions as a reference centre of science and technology for the Union. Close to the policy-making process, it serves the common interest of the Member States, while being independent of special interests, whether private or national.

JRC
EUROPEAN COMMISSION

LB-NA-22949-EN-C

Publications Office
Publications.eu.int

ISBN 978-92-79-07140-9
9 789279 071409