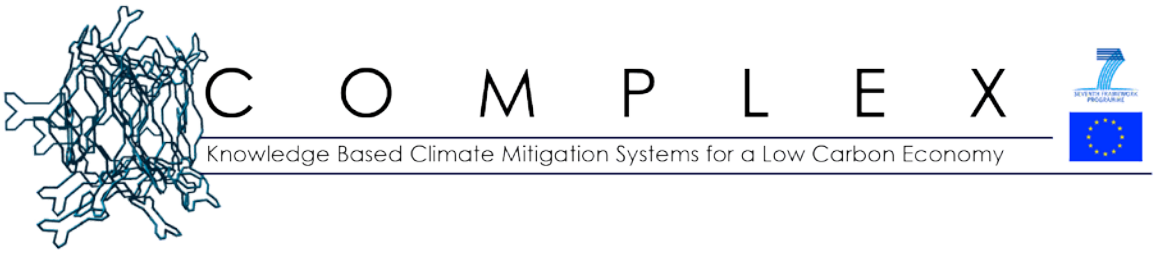

Reports on parallel modelling in the CREs and land-use case studies

Date	**Wednesday, 14 September 2016**
Report Number	**D6.9**
VERSION NUMBER:	1.0
Main Author:	Anna Shchiptsova

DIFFUSION LEVEL – PU	
PU	**PUBLIC**
RIP	**RESTRICTED INTERNAL AND PARTNERS**
RI	**RESTRICTED INTERNAL**
CO	**CONFIDENTIAL**

Coordinator: Nick Winder, nick.winder@ncl.ac.uk

INFORMATION ON THE DOCUMENT

Title	D6.9 Reports on parallel modelling in the CREs and land-use case studies
Authors	Anna Shchiptsova

DEVELOPMENT OF THE DOCUMENT

Date	**Version**	**Prepared by**	**Institution**	**Approved by**	**Note**
14.09.2016	1.0	AS	IIASA		

CONTENTS

1 Introduction

The objective of this document is to report on modeling frameworks applied in the WP6-WP3 and WP6-WP2 case studies and to describe the use of software components in conducting the analysis.

This report summarizes the following research effort:

1) The WP6-WP3 collaboration is focused on exploratory spatial analysis of regional land use patterns in the case study of Seville Province, Spain. Methodological research was conducted through Tasks 6.4 in WP6, *Development and application of data-driven stochastic qualitative models,* and required outputs from Task 3.5 in WP3.
2) The WP6-WP2 collaboration is focused on application of the posterior integration method in the case study of the sea level pressure model ensemble. Methodological research was conducted through Task 6.3 in WP6, *Development and application of methodology for integration of overlapping models,* and required outputs from Task 2.3B (i), (iii) in WP2.

2 Exploratory spatial analysis of regional land use patterns in the case study of Seville Province, Spain

2.1 Software

URL for download:
http://www.iiasa.ac.at/web/home/research/researchPrograms/AdvancedSystemsAnalysis/land-use-spatial-analysis.html

The software is supplied in jar packages.

Components:

Package 1: 'lu-preprocessing-1.0.0-standalone.jar'
Package 2: 'lu-rescaling-1.0.0-standalone.jar'
Package 3: 'lu-regression-1.0.0-standalone.jar'
Package 4: 'lu-approximation-1.0.0-standalone.jar'
Additional package: 'regression-tests-1.0.0-standalone.jar'

No installation needed. All packages are standalone java applications.

Requires JRE 1.8 installed on the target machine.

Further information can be found in COMPLEX report D6.11 (Shchiptsova, 2016).

2.2 Data

The Province of Seville is located in the Mediterranean region of Andalusia in the southwestern part of Spain. It contains the region's capital, Seville and is the largest of Andalusia's 8 provinces, both by surface area (14000 km^2) and population (1.9 m. inhabitants). A large scale, highly detailed cartographic database is freely available for this region (REDIAM, 2015). Statistical data was obtained from national government (INE, 2015; CNIG, 2015) and European government sources (EEA, 2015). The collected dataset is presented in Table 1.

Table 1. The list of collected data in the case study of the Province of Seville, Spain.

File reference	Type	Description	Year	Source
land-use.asc	GIS map	Land use classification: urban (e.g., all artificial surfaces) and non-urban (e.g., vegetation, wetlands, agricultural land and water)	2003	REDIAM, 2015
sections.asc	Panel data	Administrative division of the Province of Seville (sections)	2003	REDIAM, 2015
zoning.asc	GIS map	protected natural areas	2015	REDIAM, 2015
density.asc	Panel data	population density (people per cell)	2001	INE, 2015
distance_roads.asc	GIS map	distance to the nearest road (km)	2005	CNIG, 2015
distance_industrial_commercial.asc	GIS map	distance to the nearest area of commercial or industrial land use (km)	2006	EEA, 2015
distance_airports.asc	GIS map	distance to the nearest airport (km)	2006	EEA, 2015
distance_waterfront.asc	GIS map	distance to the nearest waterfront (km)	2005	CNIG, 2015
distance_forest.asc	GIS map	distance to the nearest area of forest (km)	2006	EEA, 2015
distance_10ths_city.asc	GIS map	proximity to a city center with more than 10,000 inhabitants (km)	2011	INE, 2015
distance_50ths_city.asc	GIS map	proximity to a city center with more than 50,000 inhabitants (km)	2011	INE, 2015

2.3 Overview

Analysis was carried as follows:

STEP 1. Preprocessing of the GIS-based variables.
The procedure includes filtering of cells with undefined values, transformation of cell values in the land use map by counting the number of cells with urban land use in the cell Moore neighborhood, logarithmic and unit rescaling transformations of the cell values in other maps, creation of spatial proximity matrix of neighboring subregions.

STEP 2. Upscaling of the GIS-based variables.
The GIS-based variables are upscaled to the level of administrative division by averaging the cell values on the GIS lattice.

STEP 3. Regression analysis with resampling.
Regression coefficients are estimated using the ordinary least squares (OLS) method. Hypothesis testing on overall model significance and individual coefficient significance is conducted using permutations method (Anderson, 2001; Freedman and Lane, 1983). Percentile bootstrap scheme (Efron, 1979; Efron and Tibshirani, 1993) is applied to estimate confidence intervals of the regression model parameters.

STEP 4. Approximation to the GIS lattice.
This step implies approximation of the GIS lattice data by the multiple linear regression model with resampling. Regression coefficients are estimated using the OLS method. Accuracy estimates are calculated for the entire GIS-based map and for the cells grouped by the number of cells with urban land use in the cell Moore neighborhood, including the cell itself. Confidence intervals for accuracy parameters are obtained using percentile bootstrap.

2.4 Preprocessing of the GIS-based variables

2.4.1 Background

The source GIS-based maps are cleaned and transformed. At first, we exclude cells, which either belong to the border of the studied area or have undefined values in their Moore neighborhood in any of the given maps in the dataset. Secondly, we drop cells, which fall into the masked areas in a "black list" GIS-based map, or cells whose neighborhood contains cells belonging to these areas.

After that, the map with land use classification is processed. In this case, we substitute every cell value in the land use map with the number of cells with urban land use in the cell Moore neighborhood.

Finally, we perform rescaling of the GIS-based maps for the specified variables. Two types of transformation are supported. Cell values on the GIS lattice can be either log rescaled or normalized.

Log rescaling is done using the formula:

$$x' = \ln(x),$$

where x is an original cell value and x' is a log rescaled value.

Normalization is applied by bringing the values of the specified variables into the range [0,1] using the following mapping:

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$

where x is an original cell value and x' is a transformed value. x_{min} and x_{max} denote the minimum and maximum values among all cells in the original GIS lattice.

2.4.2 Implementation in 'lu-preprocessing-1.0.0-standalone.jar'

Usage

$ java -jar lu-preprocessing-1.0.0-standalone.jar [options] settings-path

Arguments:

settings-path Path to the file with settings

Options:

-t, --trace Print stack trace
-h, --help Print command help

Input

;;;;; settings.xml

```
<?xml version="1.0"?>
 <files>
   <raster path = "land-use.asc" group = "land-use" />
   <raster path = "sections.asc" group = "region"/>
   <raster path = "zoning.asc" group = "mask" />

   <raster path = "density.asc" transform = "log" />
   <raster path = "distance_roads.asc" transform = "unit-rescaling " />
   <raster path = "distance_industrial_commercial.asc" transform = "unit-rescaling " />
   <raster path = "distance_airports.asc" transform = "unit-rescaling" />
   <raster path = "distance_waterfront.asc" transform = "unit-rescaling" />
   <raster path = "distance_forest.asc" transform = "unit-rescaling" />
   <raster path = "distance_10ths_city.asc" transform = "unit-rescaling" />
   <raster path = "distance_50ths_city.asc" transform = "unit-rescaling" />
</files>
```

As an input argument, 'lu-preprocessing-1.0.0-standalone.jar' receives an XML file with settings. Land use map is specified with 'land-use' value of the 'group' XML-attribute. The map should contain values equal either to 1 (e.g., all artificial surfaces) or 0 (e.g., vegetation, wetlands, agricultural land and water).

The 'region' value of the 'group' attribute defines a GIS-based map with regional administrative division. Only integer values are supported as identifiers of subregions. The 'mask' value indicates a "black list" map with mask cell values. The mask cell value should be installed to 1.

It is expected that settings xml file includes one and only one raster XML element from the 'land-use' group, one and only one raster element from the 'region' group, one and only one raster element from the 'mask' group.

Output

Results are saved to the 'lu-preprocessing' folder in the root execution directory.

Output includes cleaned and transformed .asc maps, files with Moore neighborhood statistics and a file with the matrix of spatial proximity (e.g., 'land-use.asc-neighbourhoods.csv').

2.5 Upscaling of the GIS-based variables

2.5.1 Background

The cleaned and transformed GIS-based maps are rescaled to the level of administrative division. In particular, an average of cell values in a given subregion is taken as a single observation for the GIS-based variable.

2.5.2 Implementation in 'lu-rescaling-1.0.0-standalone.jar'

Usage

```
$ java -jar lu-rescaling-1.0.0-standalone.jar [options] map-folder region-path

Arguments:
    map-folder    Path to the folder with original GIS-based maps
    region-path   Path to the file with administrative units
Options:
    -t, --trace   Print stack trace
    -h, --help    Print command help
```

Input

'map-folder' contains .asc files. E.g.,

```
"distance_roads.asc"
"'distance_industrial_commercial.asc"
"distance_airports.asc"
"distance_waterfront.asc"
"distance_forest.asc"
"distance_10ths_city.asc"
"distance_50ths_city.asc"
```

The ' region-path' argument defines a file with administrative division, e.g., 'sections.asc'.

Output

Results are saved to the 'lu-preprocessing/sample.csv' file in the root execution directory.

;;;;; 'sample.csv'

sections.a	density.as	distance_	distance_	distance_	distance_	distance_i	distance_	distance_	land_use.a
920	7.157735	0	0	0.073039	0.313935	0.03761	0	0.313949	9
558	6.326746	0	0.103834	0.259699	0.312151	0.032535	0	0.352136	9
584	-1.97735	0.336714	0.384921	0.34229	0.024488	0.244743	0.154481	0.116898	0.019694
487	5.305789	0.001239	0.351523	0.387961	0.164373	0.008174	0	0.184289	6.25
...	...	...	...	...	...	...	...	...	...

2.6 Regression analysis with resampling

2.6.1 Background

Suppose that we have data (X, y), where X is a $n \times (p+1)$ matrix of the explanatory variables and y is a $n \times 1$ vector of the response. Each column X^i is considered as the sample observations on a single explanatory variable. The sample size n is determined by the number of subregions in the studied area.

We put forward a multiple regression model in the following form

$$\begin{gathered} y = X\beta + \varepsilon \\ \varepsilon_1, \ldots, \varepsilon_n \sim F(0, \sigma^2) \end{gathered} \tag{1}$$

where $\beta = \left(\beta_0, \beta_1 \ldots, \beta_p\right)^T$ is a $(p+1) \times 1$ vector of the corresponding model coefficients respectively. By assumption, X^1 is identically 1, so that the regression equation has an intercept β_0. The error term ε is a $n \times 1$ vector of the independent identically distributed errors with common distribution F having mean 0 and finite constant variance σ^2. Both F and σ^2 are unknown to us.

For the fixed set of explanatory variables in X, the values of the unknown coefficients β are estimated using the ordinary least squares (OLS) method.

The overall significance of the specified model (1) is assessed with a permutation test with the chosen test statistic R^2 (Anderson, 2001). The significance of every individual coefficient in the specified model (1) is tested using the Freedman and Lane procedure (Freedman and Lane, 1983).

For the model (1) with the specified matrix X and unknown distribution F, the non-parametric bootstrap method (Efron, 1979) is used to measure the uncertainty associated with some prespecified model parameter on the basis of the observed data (y, X). The percentile bootstrap (Efron and Tibshirani, 1993) is applied as one of the possible bootstrap schemes. The bootstrap estimates are computed for every coefficient in β, R^2 and MSE (mean square error) parameters of the regression model.

In permutation hypothesis testing and bootstrapping, replications are done k times using the uniform generator of random permutations (with and without replacement, respectively).

To verify the absence of spatial autocorrelation in the model (1), we test whether residuals take values over distance that are more similar or less similar than expected for randomly associated pairs of observations (Overmars et al., 2003). For this purpose, the Moran's I (Moran, 1950) and Geary's C (Geary, 1954) coefficients are used

$$I = \frac{n \sum_{i,j=1\dots n} w_{ij} r_i r_j}{\sum_{i,j=1\dots n} w_{ij} \sum_{i=1\dots n} r_i^2}, \quad (6)$$

$$C = \frac{(n-1) \sum_{i,j=1\dots n} w_{ij} (r_i - r_j)^2}{2 \sum_{i,j=1\dots n} w_{ij} \sum_{i=1\dots n} r_i^2}, \quad (7)$$

where r_i is a residual value in section i and w_{ij} is an element of the matrix of spatial proximity, which is normalized by the number of neighbors of section i.

For the model (1) with unknown F, the bootstrap hypothesis test (Efron and Tibshirani, 1993) is applied with the null hypothesis of no spatial autocorrelation. The test statistic is calculated for the original data (y, X). After that, the data is resampled with replacement k times to get the reference test distribution. For every bootstrap replication, we change the matrix of spatial proximity by including repetitive observations with the same weights as the original one and adjusting the spatial weights of neighbors accordingly. The approximate p-value in the two-tailed test is expressed as (Lin et al., 2011)

$$\tilde{p}\text{-value}(\hat{\theta}) = 2 \min \left(\frac{\#\{\sigma = 1 \dots k \mid \hat{\theta}_\sigma \le \hat{\theta}\}}{k}, \frac{\#\{\sigma = 1 \dots k \mid \hat{\theta}_\sigma > \hat{\theta}\}}{k} \right), \quad (8)$$

where $\hat{\theta}$ is a test statistic of interest and $\hat{\theta}_\sigma$ is a bootstrap statistic in run σ.

After k replications, the bootstrap values $\{\hat{\theta}_\sigma\}_{\sigma=1\dots k}$ are sorted into a sequence $\hat{\theta}^*$. For the given level of confidence α, we take the $\lfloor \alpha/2\, k \rfloor$ and $\lceil (1-\alpha/2)k \rceil$ quantiles in $\hat{\theta}^*$ as the lower and upper borders of the 100×(1-α)-% percentile confidence interval. Here, $\lfloor \alpha/2\, k \rfloor$ denotes the largest integer not greater than $\alpha/2\, k$ and $\lceil (1-\alpha/2)k \rceil$ stands for the smallest integer not less than $(1-\alpha/2)k$.

2.6.2 Implementation in 'lu-regression-1.0.0-standalone.jar'

Usage

$ java -jar lu-regression-1.0.0-standalone.jar [options] path n-rep

Arguments:

path	Path to the csv file with an original sample
n-rep	Number of permutations and bootstrap replications

Options:

-t, --trace	Print stack trace
-h, --help	Print command help

Input

;;;;; 'sample-x1-x2-x3.csv'

density.asc	land_use.asc	distance_roads.	distance_indus
7.157735	9	0	0.03761
6.326746	9	0	0.032535
-1.977347	0.019694	0.154481	0.244743
5.305789	6.25	0	0.008174
...	...	...	...

The 'path' argument defines a file with sample values, e.g., 'sample-x1-x2-x3.csv'. It is expected that the first row contains variable labels. The first column should contain values of the response y.

Output

Results are saved to the 'lu-regression' folder in the root execution directory. The 'lu-regression/ permutation_tests.csv' file contains results of permutation testing. Bootstrap estimates are saved to 'lu-regression/regression-stat-bootstrap.csv'.

;;;;; 'permutation_tests.csv'

test	p-value	lower-bound-ci	upper-bound-ci
overall-test-r2	0.0001	-0.000096	0.000296
land_use.asc-test-t-stat	0	0	0
distance_roads.asc-test-t-stat	0	0	0
distance_industrial_commercial.asc-test-t-stat	0.007699	0.005986	0.009412

;;;;; 'regression-stat-bootstrap.csv'

statistics	95-percent-ci-1	95-percent-ci-2	mean
land_use.asc	0.370039	0.453511	0.414008
distance_roads.asc	-45.80548	-25.320336	-34.475848
distance_industrial_commercial.asc	-5.897927	-1.483479	-3.698972
intercept	2.658128	3.310021	2.967782
r-squared	0.829551	0.872105	0.85143
mse	0.651385	0.86007	0.753846

2.6.3 Implementation of spatial autocorrelation test in 'regression-tests-1.0.0-standalone.jar'

Usage

$ java -jar regression-tests-1.0.0-standalone.jar [options] path n-replications "iid" path2

Arguments:

path	Path to the csv file with sample data
n-replications	Number of replications in permutations and bootstrapping
path2	Path to the additional csv file

Options:

-t, --trace Print stack trace

-h, --help Print command help

Input

;;;;; 'sample-x1-x2-x3.csv'

density.asc	land_use.asc	distance_roads.	distance_indus
7.157735	9	0	0.03761
6.326746	9	0	0.032535
-1.977347	0.019694	0.154481	0.244743
5.305789	6.25	0	0.008174
...	...	...	...

;;;;; 'sections-neighbours.csv'

id	n1	n2	n3	n4	n5	n6	n7	...
1	2	361	383	535	541			...
2	1							...
3	4	391						
4	3							...
5	531	534	669					...
7	9	13	16					...
9	7	10	11	13				...
10	9	11	13	14	29	43		...
...	...	...	...	...	...	...	...	...

The 'path' argument defines a file with sample values, e.g., 'sample-x1-x2-x3.csv'. It is expected that the first row contains variable labels. The first column should contain values of the response y.

The 'path2' argument defines a file with the matrix of spatial proximity. It is expected that the first column contains identifiers of administrative units, e.g., ids of sections. The row values contain identifiers of the neighboring administrative units (e.g., ids of sections) to the current unit.

Output

Results are saved to the 'regression-tests' folder in the root execution directory. The 'regression-tests/independence-tests-bootstrap.csv' file contains results of bootstrap hypothesis testing. The generated bootstrap samples are saved to 'regression-tests/morans-i-test-sample.csv' and 'regression-tests/geary-c-test-sample.csv'.

;;;;; 'independence-tests-bootstrap.csv'

statistics	95-percent-ci-1	95-percent-ci-2	mean	p-value
morans-i-test	-0.074169	0.085302	0.005006	0.689131
geary-c-test	0.850017	1.143669	0.993253	0.972303

;;;;; 'morans-i-test-sample.csv'

value
0.016686
-0.02685
0.024039
0.010429
...

;;;;; 'geary-c-test-sample.csv'

value
0.944214
1.10823
0.969776
0.944674
...

2.7 Approximation to the GIS lattice

2.7.1 Background

Each subregion matches some area in the GIS-based map represented by a regular lattice of cells. We consider model (1) as a stochastic approximation to the true response value in every cell of the map. Suppose that we have data $(X', \tilde{y}')$, where X' is a $N \times (p+1)$ matrix of values of the explanatory variables and $\tilde{y}'$ is a $N \times 1$ vector of the true observed values at the cell level. We denote by y' a $N \times 1$ vector of the stochastic response values in (1), corresponding to the given explanatory variables X'. N is a total number of cells in the map.

For the given cell j, we define an accuracy ρ_j of the response value y'_j from (1) to the true value $\tilde{y}'_j$, observed in this cell, as a distance between the expected model response and the true value. That is

$$\rho_j = \left|\tilde{y}'_j - \hat{y}'_j\right|, \tag{2}$$

where $\hat{y}'_j = X'_j\beta$ and $j = 1, \dots, N$. $\hat{y}'_j$ is the fitted value of the response in (1) for the observed values $X'_j = \left(x'_{j1}, \dots, x'_{jp}\right)$ of the explanatory variables in the cell j. Here, we enumerate cells in the GIS-based map, and denote by j an index of the cell in this enumeration.

We map approximation by the model (1) to a value of the accuracy ρ_j in every cell. Therefore, a sample of the accuracy values is obtained for the GIS-based map. For the given subset of cells S, we measure the $(100 \times k)$-th percentile of the accuracy of the response values $\{y'_j\}_{j \in S}$ from (1) to the true values $\{\tilde{y}'_j\}_{j \in S}$, observed in the cells belonging to this subset, as a minimum value below which at least the k-th fraction of the cell accuracy values fall. That is

$$\rho(k, S) = \underset{\bar{\rho} \geq 0}{\operatorname{argmin}}[\#\{\rho_i \leq \bar{\rho} \mid i \in S\} \geq k \cdot m], \tag{3}$$

where $0 \leq k \leq 1$, S is a subset of the cell indexes with m elements. We denote by # the number of elements in the set. The value $\rho(1, S)$ matches the maximum accuracy in the given subset of cells. The quartiles in the obtained sample of accuracy values for the cells in S coincide with the values of $\rho(0.25, S)$, $\rho(0.5, S)$ and $\rho(0.75, S)$. The accuracy estimates are denoted by $\rho_{max}(S)$, $\rho(Q_1, S)$, $\rho(Q_2, S)$ and $\rho(Q_3, S)$ respectively.

To apply the formula (2), we bootstrap the data (y, X) to get the reference distribution for the accuracy statistic. In particular, bootstrap estimates for the maximum accuracy and for the accuracy quartiles for the selected sample of ρ_j are obtained. Specifically, we project an estimated vector of coefficients $\hat{\beta}_\sigma$ in the model (1) to the map accuracy values and summarize the sample by the percentiles in every bootstrap replication. Thus, separate bootstrap samples for $\rho_{max}(S)$, $\rho(Q_1, S)$, $\rho(Q_2, S)$ and $\rho(Q_3, S)$ are found, and percentile confidence intervals are calculated for these parameters.

2.7.2 Implementation in 'lu-approximation-1.0.0-standalone.jar'

Usage

$ java -jar lu-approximation-1.0.0-standalone.jar [options] sample-path values-path n-rep

Arguments:

sample-path	Path to the csv file with an original sample
values-path	Path to the csv file with target values for approximation
n-rep	Number of bootstrap replications

Options:

-t, --trace	Print stack trace
-h, --help	Print command help

Input

;;;;; 'sample-x1-x2-x3.csv'

density.asc	land_use.asc	distance_roads.	distance_indus
7.157735	9	0	0.03761
6.326746	9	0	0.032535
-1.977347	0.019694	0.154481	0.244743
5.305789	6.25	0	0.008174
...	...	...	...

;;;;; 'cells-x1-x2-x3.csv'

density.as	group	land_use.	distance_	distance_i
-1.69062	0	0	0.406045	0.550763
-1.69062	0	0	0.416017	0.555664
-1.69062	0	0	0.42604	0.560662
-1.69062	0	0	0.436063	0.565755
...	...	...	...	...

The 'sample-path' argument defines a file with original sample values, e.g., 'sample-x1-x2-x3.csv'. It is expected that the first row contains variable labels. The first column should contain values of the response y.

The 'values-path' argument defines a file with target values, e.g., 'cells-x1-x2-x3.csv'. It is expected that the first row contains variable labels. The first column should contain values of the response y. The second column contains the group id for the given cell value.

Output

Results are saved to the 'lu-approximation/accuracy-bootstrap.csv' file in the root execution directory.

;;;;; 'accuracy-bootstrap.csv'

id	group-id	95-percent-ci-1	95-percent-ci-2	mean
p-25-percent-0	0	0.91972	1.154213	1.034962
p-25-percent-1	1	1.648701	1.834848	1.740372
p-25-percent-all	all	0.949766	1.181227	1.063125
p-50-percent-0	0	1.802652	2.270311	2.02227
p-50-percent-1	1	3.189866	3.481749	3.337134
p-50-percent-all	all	1.875895	2.342561	2.094395
p-75-percent-0	0	2.844491	3.490305	3.134313
p-75-percent-1	1	4.402656	4.712864	4.555963
p-75-percent-all	all	2.976079	3.602863	3.256569
p-max-0	0	11.925989	20.971695	15.92601
p-max-1	1	9.958119	14.244063	10.7058
p-max-all	all	11.925989	20.971695	15.92626
p-min-0	0	0.000001	0.000076	0.00002
p-min-1	1	0.000009	0.001301	0.000361
p-min-all	all	0	0.000072	0.000019

2.8 Results

Table 2. Approximate p-values with 95% normal approximation confidence intervals. Model specification: X_1 - number of urban cells in the Moore neighborhood, X_2 - distance to a road, X_3 - distance to a commercial center (number of replications 10,000).

statistics	$\tilde{p}$-value
t_1	0.000 (0.000, 0.000)
t_2	0.000 (0.000, 0.000)
t_3	0.008 (0.006, 0.010)
R^2	0.000 (0.000, 0.000)

Table 3. Bootstrap estimates. Model specification: X_1 - number of urban cells in the Moore neighborhood, X_2 - distance to a road, X_3 - distance to a commercial center (number of replications 10,000)

Coefficient	Mean	95% percentile confidence interval	Statistic	Mean	95% percentile confidence interval
β_1	0.414	(0.370, 0.454)	R^2	0.851	(0.829, 0.872)
β_2	-34.415	(-45.704, -25.389)	MSE (mean square error)	0.754	(0.654, 0.859)
β_3	-3.700	(-5.889, -1.416)			
β_0	2.966	(2.656, 3.301)			

Table 4. Bootstrap estimates and approximate p-values for the measures of spatial autocorrelation. Model specification: X_1 - number of urban cells in the Moore neighborhood, X_2 - distance to a road, X_3 - distance to a commercial center (number of replications 10,000).

Statistic	Mean	95% percentile confidence interval	p-value
Moran's I	0.005	(-0.074, 0.085)	0.689
Geary's C	1.051	(0.85, 1.144)	0.972

Table 5. Bootstrap mean with 95% percentile confidence intervals for the accuracy percentiles in the cells grouped by the number of cells with urban land use in the cell Moore neighborhood. Model specification: X_1 - number of urban cells in the Moore neighborhood, X_2 - distance to a road, X_3 - distance to a commercial center (number of replications 10,000). "All": the sample of the entire GIS-based map, "non-urban": the sample of cells with no urban cells in the Moore neighborhood, "urban": the sample of cells with at least one urban cell in the Moore neighborhood (including the cell itself).

S	number of points	$\rho(Q_1, S)$	$\rho(Q_2, S)$	$\rho(Q_3, S)$	$\rho_{max}(S)$
Group "non-urban"	216,729	1.04 (0.92, 1.15)	2.02 (1.81, 2.29)	3.14 (2.85, 3.50)	15.95 (11.92, 20.97)
Group "urban"	18,949	1.74 (1.65, 1.84)	3.34 (3.20, 3.48)	4.56 (4.41, 4.71)	10.72 (9.96, 14.18)
Group "all"	235,678	1.06 (0.95, 1.18)	2.10 (1.88, 2.34)	3.26 (3.00, 3.61)	15.95 (11.92, 20.96)

3 Posterior integration in the case study of the sea level pressure model ensemble

3.1 Software

URL for download:
http://www.iiasa.ac.at/web/home/research/researchPrograms/AdvancedSystemsAnalysis/modelIntegration-package.html

The software is supplied as R package 'modelIntegration'.

Either a Windows binary package 'modelIntegration_1.0.0.zip' or a bundled package 'modelIntegration_1.0.0.tar.gz' can be downloaded.

On a Windows platform only:
install a binary package
``` R
install.packages(path_to_folder/modelIntegration_1.0.0.zip, repos = NULL)
```
where 'path_to_folder' will represent the full path to the local directory.

On all platforms:
install from a source distribution (build tools should be installed)
``` R
install.packages(path_to_folder/modelIntegration_1.0.0.tar.gz, repos = NULL, type = "source")
```
where 'path_to_folder' will represent the full path to the local directory.
On Windows it will look something like this:
"C:\\download folder\modelIntegration_1.0.0.tar.gz".
On UNIX it will look like this: "/home/download/modelIntegration_1.0.0.tar.gz".

Further information can be found in COMPLEX report D6.10 (Shchiptsova, 2016).

3.2 Data

We use seasonal sea level pressure (SLP) time series observed at longitude 5W, latitude 44N in period 1871 – 2004 (134 years) and four corresponding seasonal SLP time series simulated by the ensemble of models 1 – 4 for the same period and region. Totally, the posterior integration method was executed in 12 runs (3 runs per season).

The following list of abbreviations is used:

- Months
  - JFM (January, February, March)
  - AMJ (April, May, June)
  - JAS (July, August, September)
  - OND (October, November, December)

- Periods
  - Run 1 (analysis period 1871–1915, 1916–1959; test period 1960–2004)
  - Run 2 (analysis period 1871–1915, 1960–2004; test period 1916–1959)
  - Run 3 (analysis period 1960–2004, 1916–1959; test period 1871–1915)

- Models
  - Data (data-based seasonal SLP probability distribution)
  - Model 1–4 (seasonal SLP probability distributions based on models 1-4)
  - Product distribution (Kryazhimskiy, 2013; Kryazhimskiy, 2016)

3.3 Theoretical background

Suppose that, several independent methods are used to observe a deterministic element and each method represents the latter as a probability distribution. Thus, we deal with a family of probability distributions providing alternative descriptions to the same object. The problem is how to combine information from the prior estimates.

The posterior integration method (Kryazhimskiy, 2013; Kryazhimskiy, 2016) is based on the assumption that model outcomes are mutually compatible, i.e., we should observe identical outcomes after the use of model ensemble. Formally, the product probability distribution of the original estimates is

$$p(z) = \frac{p_1(z) * p_2(z) * \ldots * p_n(z)}{\sum_{z' \in Z} p_1(z') * p_2(z') * \ldots * p_n(z')},$$

where $p_1, p_2, \ldots, p_n$ are prior distributions on Z associated with the methods $1, \ldots, n$. Z is a non-empty finite set, whose number of elements is bigger than one.

3.4 Implementation in the 'modelIntegration' R package

Usage

```
> library(modelIntegration)
> data <- read.csv("JFM_Run1.csv")

> integration <- integrate(data[, 1], as.list(data[, 2:5]))
> statistics(integration)
##                M1          M2          M3          M4       Product
## mean 101628.5335 101860.9867 101416.7429 101672.4414 101543.9325
## std     368.6282    416.1136    228.2541    312.8084    164.6561
##              Average
## mean  101644.6761
## std      373.7518
```

```
> print(integration)
##         x       M1         M2         M3         M4         Product
## 1  100683.2 0.00000000 0.00000000 0.00000000 0.01123596 0.0000000000
## 2  100913.1 0.04494382 0.03370787 0.03370787 0.01123596 0.0001027611
## 3  101143.0 0.15730337 0.02247191 0.20224719 0.02247191 0.0028773118
## 4  101372.8 0.14606742 0.16853933 0.40449438 0.22471910 0.4007684249
## 5  101602.7 0.22471910 0.12359551 0.26966292 0.33707865 0.4521489922
## 6  101832.6 0.21348315 0.21348315 0.07865169 0.21348315 0.1370519543
## 7  102062.4 0.14606742 0.21348315 0.01123596 0.11235955 0.0070505556
## 8  102292.3 0.06741573 0.11235955 0.00000000 0.05617978 0.0000000000
## 9  102522.2 0.00000000 0.10112360 0.00000000 0.01123596 0.0000000000
## 10 102752.1 0.00000000 0.01123596 0.00000000 0.00000000 0.0000000000
```

3.5 Results

Table 6. Models' performance in cross-comparison runs, season JFM.

Model	Run 1		Run 2		Run 3	
	mean	std	mean	std	mean	std
Model 1	101371.241	223.26	101376.101	221.44	101347.448	194.18
Model 2	101991.702	235.26	102011.388	244.41	102020.986	265.71
Model 3	101415.167	204.02	101419.540	198.31	101431.640	202.24
Model 4	101686.047	316.77	101687.410	315.27	101746.446	272.47
Product	101650.474	130.95	101586.901	175.46	101580.188	149.54

Table 7. Models' performance in cross-comparison runs, season AMJ.

Model	Run 1		Run 2		Run 3	
	mean	std	mean	std	mean	std
Model 1	101426.368	147.41	101387.434	144.80	101428.692	140.91
Model 2	101472.841	160.30	101505.773	160.31	101508.858	177.14
Model 3	101264.874	132.23	101274.840	131.68	101285.787	129.77
Model 4	101504.211	195.61	101496.581	161.70	101503.049	178.35
Product	101391.621	75.63	101415.802	67.29	101387.924	79.54

Table 8. Models' performance in cross-comparison runs, season JAS.

Model	Run 1		Run 2		Run 3	
	mean	std	mean	std	mean	std
Model 1	101571.413	90.38	101594.403	96.29	101587.354	92.92
Model 2	101705.318	105.80	101718.939	116.63	101726.042	105.00
Model 3	101506.054	90.63	101498.243	84.07	101509.242	84.35
Model 4	101879.077	135.45	101878.156	152.85	101882.265	137.21

Product	101622.365	47.34	101609.504	43.77	101634.961	53.13

Table 9. Models' performance in cross-comparison runs, season OND.

Model	Run 1		Run 2		Run 3	
	mean	std	mean	std	mean	std
Model 1	101371.241	223.26	101376.101	221.44	101347.448	194.18
Model 2	101991.702	235.26	102011.388	244.41	102020.986	265.71
Model 3	101415.167	204.02	101419.540	198.31	101431.640	202.24
Model 4	101686.047	316.77	101687.410	315.27	101746.446	272.47
Product	101650.474	130.95	101586.901	175.46	101580.188	149.54

References

[1] Anderson, M. (2001). Permutation tests for univariate or multivariate analysis of variance and regression. Canadian Journal of Fisheries and Aquatic Sciences, 58(3): 626-639. DOI: 10.1139/f01-004

[2] Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. Annals of Statistics, 7(1): 1-26. DOI:10.1214/aos/1176344552

[3] Efron, B., & Tibshirani, R. (1993). An Introduction to the Bootstrap. New York: Chapman and Hall.

[4] Freedman, D., & Lane, D. (1983). A Nonstochastic Interpretation of Reported Significance Levels. Journal of Business & Economic Statistics, 1(4): 292-298. DOI: 10.2307/1391660

[5] Geary, R. (1954). The Contiguity Ratio and Statistical Mapping. The Incorporated Statistician, 5(3): 115-145. DOI: 10.2307/2986645

[6] Kryazhimskiy, A.V. (2013). Posterior integration of independent stochastic estimates. IIASA Interim Report. IR-13-006.

[7] Kryazhimskiy, A.V. (2016). A posteriori integration of probabilities. Elementary theory. Theory of Probability and its Applications, 60(1): 62-87. DOI: 10.1137/S0040585X97T987466

[8] Lin, K.-P., Long, Z.-H., & Ou, B. (2011). The Size and Power of Bootstrap Tests for Spatial Dependence in a Linear Regression Model. Computational Economics, 38(2): 153-171. DOI: 10.1007/s10614-010-9224-0

[9] Moran, P. (1950). Notes on Continuous Stochastic Phenomena. Biometrika, 37(1-2): 17-23. DOI: 10.2307/2332142

[10] Overmars, K., de Koning, G., & Veldkamp, A. (2003). Spatial Autocorrelation in Multi-scale Land Use Models. Ecological Modelling, 164: 257-270. DOI: 10.1016/S0304-3800(03)00070-X

[11] Shchiptsova, A. (2016). Software package 1: integration of models. COMPLEX project report D6.10.

[12] Shchiptsova, A. (2016). Software package 2: data-driven stochastic modelling. COMPLEX project report D6.11.

Web References

[1] CNIG (2015). Download Center of the National Center for Geographic Information. http://centrodedescargas.cnig.es/CentroDescargas/inicio.do/. Accessed 1.08.16.

[2] EEA (2015). European Environment Agency, Copernicus Land Monitoring service CORINE land cover. http://land.copernicus.eu/pan-european/corine-land-cover/. Accessed 1.08.16.

[3] INE (2015) Instituto Nacional de Estadística. http://www.ine.es/. Accessed 1.08.16.

[4] REDIAM (2015) Andalusian Government environmental information service. http://www.juntadeandalucia.es/medioambiente/site/rediam/. Accessed 1.08.16.