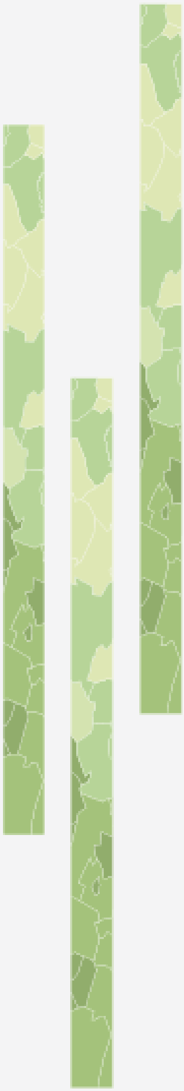
TECHNICAL REPORT

JUNE 2012



Detecting and Handling Anomalous Data in M4D

Combined Report

CONTENT

Anomalous data represent a challenge both for suppliers and users of M4D data

Anomalous data can arise from errors in the data handling and management process

Anomalous data can also arise naturally and is a characteristic of unusual areas

Accurate identification of anomalous data is import: errors need to be handled appropriately before the data are loaded into the Database

Errors can sometimes be identified mechanically as well as statistically. True outliers are identified statistically

Anomaly detection approaches depend on the type of data being checked

An approach has been implemented using the R open source software

38 pages

ESPON M4D -
MULTI DIMENSIONAL DATABASE DESIGN & DEVELOPMENT

EUROPEAN UNION
Part-financed by the European Regional Development Fund
INVESTING IN YOUR FUTURE

LIST OF AUTHORS

Martin Charlton, NCG

Paul Harris, NCG

Alberto Caimo, NCG

Contact

martin.charlton@nuim.ie

National Centre for Geocomputation
National University of Ireland Maynooth
Maynooth
Co Kildare
IRELAND
Tel: (+353) 1 708 6186

TABLE OF CONTENTS

Introduction

An important aspect of the ESPON Database 2013 Phase II project (hereinafter referred to as M4D[1]) is the explicit attention paid to the detection and appropriate handling of potentially anomalous data. A set of anomaly detection techniques suitable for M4D data has been developed by a team at the National Centre for Geocomputation. This builds on existing ideas for techniques developed in the earlier project: ESPON Database 2013, and part of this document is based on the techniques described in the report: Spatial Analysis for Quality Control Phase 1: the identification of logical input errors and statistical outliers.

A major function of the M4D Database is to act as a repository for data created and used by the various project under the ESPON programme. The data typically are supplied towards the end of the project period. The data flow, checks, and processes which form the workflow from supplier to database have been developed by the M4D partners. Notable is the importance given to the metadata for every indicator in each dataset. This is used in the data quality check process to determine which tests are to be applied to the dataset in order to build the weight of evidence in favour of an observation being unusual or not.

This combined report covers **three** of the NCG deliverables – (i) methods for spatially normalised data, (ii) methods for spatially un-normalised data and (iii) consideration of an implementation of the methods . Reports (i) and (ii) are really two components of something that is logically connected. This leadss into (iii) which appraoch is described in detail and is illustrated with a breif example session.

The overall approach is driven by

(a) the type of the data (typology, stock, ratio…)
(b) the spatial units (NUTS, UMZ…)
(c) the application of exploratory techniques
(d) application of tests taking indicators singly, in pairs and in groups
(e) the application of aspatial and spatial tests
(f) the organisation of the datasets from the ESPON projects
(g) the timetable for the dataflow.

There is still some minor development work to undertake which will be completed shortly.

[1] M4D: **M**ulti**D**imensional **D**atabase **D**esign and **D**evelopment

Outlier checking and the Database

Work Package A2 deals with the problems which are attendant on the presence or otherwise of anomalies in the data for the Database. In the Inception report the three subsections of this package are Time Series, Outlier Handling, and Data Quality. As work on these has progressed it is clear that some adjustment to the original somewhat vague intentions is required, and that these intentions need to be made somewhat clearer.

In their wide ranging survey Chandola et al (2007) assert that "Anomalies are patterns in data that do not conform to a well-defined notion of normal behaviour". They also point out that such notions are often domain dependent, and this is reflected in the terminological variation which is encountered; this includes: anomaly, outlier, discordant observation, aberration, surprise, peculiarity, and contaminant. There are a wide variety of anomaly detection approaches which have been developed over the years, but many of these are specific to particular domains, such as the detection of credit card fraud. They suggest that there are several challenges in anomaly detection:

defining a 'normal' region is difficult

normal behaviour tends to evolve

the notion of an anomaly is different in different application domains

the availability of training ('labelled') data is a major issue

the noise inherent in the data may make anomalies difficult to detect

Hawkins (1980) opens his monograph with the pragmatic definition of an outlier to be "an observation which deviates so much from other observations as to arouse suspicion that it was generated by a different mechanism". Another view of outliers is that they are one or more observations that greatly influence the value of a statistical estimator. Chandola in general and Hawkins in particular conceive of an outlier as something that arises from some process, whereas we might also think of an outlier that has a particular outcome on the result of an experiment.

Earlier work on the identification of outliers includes Pearson and Sekar (1936); in the 1950s researchers such as Grubbs (1950, 1951) and Dixon (1950) developed simple tests for detection. Hawkins (1980) summarised the state of the art in his survey of identification techniques. More recently the development of computational data mining techniques has led further work. Ben-Gal, (2005) summarises recent developments and provide an overview. Most recent work includes that of Peter Filzmoser and his colleagues at Technische Universität Wien in Filzmoser *et al* (2005, 2008, 2012).

Software libraries for outlier detection include three R packages (available at http://cran.r-proejct.org). Komsta's *outliers* package implements some simple univariate tests including those of Grubbs, Dixon and Cochran. Van der Loos' package extremevalues provides an interactive approach to the identification of extreme values in one-dimensional data. Peter Filzmoser is a leading researcher in this area, and has contributed mvoutlier, a package which provides methods for outlier detection in multivariate data based on robust methods.

Ben-Gal (2005) divides outlier detection tests into those *univariate* data and those for *multivariate* data. Earlier work in outlier detection concentrated on a single indicator whereas considering the behaviour of several indicators simultaneously has been the subject of more recent research. Methods may also be *parametric* (that is, they assume an underlying model for the data, often the normal distribution) or *non-parametric* (they do not assume the data arise from a particular model). Techniques may also be *distance based* – they are based on lcaol distance measures or *cluster based* – they attempt to detect clustered outliers. Finally we should also note that there exist tests for spatial outliers. An interesting phenomenon is that an observation which may be an outlier in a multivariate sense (the combination of indicators is unusual), it need not be an outlier in the univariate case.

It is clear that a single general purpose anomaly detector into which an ESPON project dataset can be placed and which prints a report highlighting the NUTS regions in which there appear to be anomalous data is perhaps a chimera. The range of ESPON projects, and the diversity of the data which is offered as potential input to the database mean that we have to think of a range of generic techniques which can be selected as appropriate and applied to a particular dataset. It is also desirable that the anomaly detection that is run on a dataset could be re-run at a future date, and which would yield the same results: in other words, the process should be reproducible.

We need to make some simplifying assumptions about the potential uses of the Database data; what may be uncomfortable data for one application may be that which highlights special cases in another. A 'one-size-fits-all' rule is not appropriate. However, data that is clearly *wrong* should be corrected. Identifying that which is wrong is context specific. For example an indicator which is a categorisation with 4 categories which the metadata list as 1,2, 3, 4 should contain no other values apart from those; data which are counts should belong to the set of positive integers – we cannot have a negative count of residential population. More challenging are indicators such as the male unemployment rate; some regions with particularly pressing economic problems may have very high unemployment rates which lie far in the upper tail of the distribution of values. Income data tend to reflect the factor that most people in a region have modest incomes, but there are a few very rich individuals among them. Whilst such upper tail values may be anomalous, they are not wrong.

We might then consider the nature of the methods we use to detect anomalies, and whether they are influenced by the anomalies themselves. Wright and London (2009) show that statistics which make assumptions about the underlying distribution of the data (for example: that it follows a normal distribution) may be unduly influenced by anomalies. The presence of anomalous values tends to lower the chance of finding a significant result in a statistical analysis.

ESPON datasets and data quality checking

The data sent from the ESPON projects has a well-defined layout. Each spreadsheet contains four labelled worksheets, containing (a) general information about the dataset (b) metadata about the indicators (c) metadata on the sources and (d) the data itself. The data tables are organised such that rows represent observations (usually NUTS regions) and the columns represent indicators. The columns are logically organised into pairs such that the first contains the data values and the second contains the source key.

We have chosen to use the R statistical environment (R, 20xx) for creating the functions that undertake the anomaly detection. First, R is free and open source software; the anomaly detection is not tied to any one commercial product. Second, there is an enormous range of contributed packages which implement an equally enormous range of statistical methods. Third, users can create their own functions, created from different combinations of the existing functions, rapidly and easily. Fourth, R scripts provide a reproducible approach to the analysis of data. A function in R can range from a simple function that returns the sine of a given angle, through ones that fit complex multivariate models, to ones that read and act on spatial data, and finally functions that summarise and display in a range of visual formats. R's strength is its flexibility.

Our approach has been to provide a *library* of detection methods which can be assembled in a series of templates. The templates provide a basic outline for the different sorts of data which we expect to encounter in the data quality activities. Currently there are two templates, one for time-series data, and one for cross-sectional data. An *instance* of a template is the result of its adaption for a particular dataset. Eventually there will be at least as many instances as there are ESPON datasets supplied for data quality assessment.

The output of data for the Database by an ESPON project is in the form of one or more Excel spreadsheets. We have adopted the principle that we do not work directly on the spreadsheet: rather, we copy the data into R and work on that copy. The copying is done invisibly and does not require human agency save that of invoking three R functions.

The following sections outline (a) the issues and approaches involved in handling time series (b) the issue and approaches involved in handling cross-sectional data and (c) the modus operandi for dealing with the data quality operation in the M4D workflow.

Outlier detection

We start by considering the nature of the data in each indicator. A binary classification of data is that the values for either categorical or continuous. Categorical data have the property that a small number of distinct values exist: such data can be either nominal or ordinal. That is: either the categories are unordered, or the categories are ordered. We often refer to the latter as ranks. Continuous data are sometimes categorised into interval data and ratio data. Interval data are those in which there is no well-defined zero (for example: the zero points in the Fahrenheit and Celsius temperature scales are arbitrary). Ratio data has well defined zeros, such that the ratio of two value has a sensible interpretation.

We have an additional problem with spatial data. The indicators form the attributes of the spatial units which have been used in the work undertaken by the Project Partner who has supplied the dataset for checking. The tests we can conveniently use require that the spatial data are consistent, and that any regions with missing values have been removed. Islands present an interesting problem too: they have no adjacent neighbours so the computation of the spatial weights matrices for the spatial tests should have cognisance of this.

In examining the datasets for anomalous data there are two main approaches: exploratory and statistical. The exploratory approaches we have comprise initial data summaries, and exploratory visualisations, and the latter lead us towards creating evidence in favour of an observation have one or more items of anomalous data.

Logical input errors

Logical input errors can arise for a number of reasons. For example, the wrong NUTS1 code could be specified; incorrect data values could be input; data could be repeated exactly but assigned to different variables; data could be displaced within or between columns; data could be swapped within or between columns. In general, the identification of an input error will follow some logical, mathematical approach. For example, if a land use class could only take a positive integer value from 1 to 9 say, then an input error of say, -2, 4.5 or 10 would be easily identified.

An input error may also be identified statistically. For example, if the number 27 is inadvertently entered as 72 for a region‟s unemployment rate, the value 72 may lie in the extreme tail of this variable‟s distribution and as such, is statistically-outlying. A difficulty here would be to distinguish between an input error of 72 and a true value of 72.

In this respect, when dealing with errors/outliers, most input errors can be either be corrected or removed, whilst most outliers should be flagged as: (i) suspected outliers and (ii) potential (undetected) input errors. Flagged observations would then require further scrutiny, which should ascertain whether the observation should be: (a) replaced; (b) removed; or if specifically an outlier, (c) retained or possibly down-weighted in some way (so as to provide some robust model fit or statistic of the data).

Exploratory approaches

The nature of the initial exploration depends on the type of data that we have. Initially we will consider each indicator in the dataset singly. For **nominal** data a *frequency tabulation* will show whether there are values outside the allowable set presented in the metadata; a *barplot* provides a quick visual summary of the numbers of regions in each category. For **count** data a *five number summary* which shows the upper and lower extremes, quartiles and the median will reveal whether the data are all positive; a *histogram* of the values will also give a visual indication of this as well.

For **ratio** data a *five number summary* gives a brief overview of the distribution, and a *boxplot* will give an initial visualisation of the likelihood that any anomalous values are present in the data. For **time series** data, a parallel coordinates plot provides an initial visualisation of the general patterns in the data.

The data for the atlas relates to spatial units, so mapping the data is an obvious visualisation of the spatial patterns in the data. In an initial exploration of the spatial properties of the data quick visual summaries of the indicators are possible for *nominal* and *ratio* data. For *count* and *time series* data, the spatial pattern is also conditioned on the underlying population at risk – as this is the likely not to be known, a proxy might take the form of *area-normalising* (dividing the counts by the area in square kilometres of their respective spatial units). Following the suggestions of Tufte (1983) we avoid the use of pie charts, or collections of pie charts which require the user to make comparisons both with and between the pies. It should be noted here that such graphics are essentially ephemeral and exploratory – they need not conform to ESPON layout specifications.

Bivariate and multivariate displays take use into more complex realms which are domain specific. That we can compute correlations or plot scatterplot matrices does not necessarily have the outcome that the resulting displays are meaningful in the context of the work undertaken in the Project, or helpful. We have seen some data expressed as principal component scores – a correlation matrix will reveal whether the components are orthogonal (the correlations should all be close to zero across the set of components). We have also seen regression residuals – these should be independent, centred on zero heteroskedastic, and should not exhibit non-zero spatial autocorrelation. However, regression residuals are not currently candidates for inclusion in the database, so they will be omitted from outlier checks.

Statistical approaches

We can divide the tests for outlier into those that are aspatial (the physical arrangement of the spatial units is not relevant) and those that are spatial (they take cognisance of the spatial arrangement of underlying areas).

The tests to be applied depend on the nature of the indicator in question.

Nominal data

Nominal data consists of a number of distinct categories. The metadata contains the list of valid codes used to represent the separate categories. A simple aspatial test is to consider whether any of the entries in the indicator have values which are not present in the list of valid codes. The %in% function in R will return TRUE is an individual value is in the list, FALSE if not. If indicator a vector containing the list of codes to be tested, and valid.code is a vector of valid codes, then the R statement:

```
test <- indicator %in% valid.codes
```

… will create a vector, test, whose FALSE entries indicate entries in indicator with invalid codes. A list of the NUTS regions which contain out of range data can be created; assuming UnitCode is the vector of NUTS codes for the dataset undergoing checking, then UnitCode[!test] will be a list of the regions for which the indicator being tested has out of range data.

A second test is to consider the entries in a table showing the numbers of regions in each category of the indicator. If there were four categories, and an equal spread of 100 regions in each category it would be difficult to suggest that any category was anomalous. However, if there were 33 each in three categories and 1 in the fourth, then that category might well be regarded as unusual. In this case Grubbs test (Grubbs, 1950) can be employed to determine whether this category can be regarded as unusually sparsely represented; a second version of Grubbs test can be used to determine whether there are two unusually low frequency categories in the data.

Spatial tests for outliers, such as a Local Moran's I test, assume that the data under test is continuous and is reasonably normally distributed. With nominal data this is not the case, and we are restricted to simple tests against the metadata.

Ordinal Data

Ordinal data are sometimes known as ranks – the values indicate the position of an observation relative to other observations on an ordered scale where the difference between adjacent values is unity. If there are N regions, then we would expect that the ranks would run from 1 to N. However, they would not necessarily belong to the set of positive integers, as zoned with tied ranks may be allocated an interpolated value: for example, if two observations both had the highest value of a variable, the resulting rank might be entered as 1.5 {(1+2)/2}. As these data are not ratio scale (that is, the ratio of the top two values is different from the ratio of the next two (1/2 compared with 3/4) these data do not lend themselves to tests for ratio data. The range of values can be checked: they should be greater than or equal to 1 and not greater than N. The R code would be:

```
test <- indicator >= 1 & indicator <= N
```

The vector test would contain TRUE for those observations whose values fell within the desired value range, and FALSE otherwise. Any observations for whom the filter was FALSE may be in error on the tested indicator.

If the ordinal data one or more indicators can be treated as a metric (there are certain conditions for the existence of a metric) then we may be able to apply some of the ratio based methods below.

Ratio data: unnormalised

Count data for spatial units provides some interesting challenges. Count data are either zero or a positive integer. Count data are ratio data, but when they are presented for spatial units, the differing sizes of the spatial units means that the values cannot be treated equally. Consider population as an example; if the population density in the NUTS regions was uniform, then the values of the population count would be in proportion to the area of the region. However, this is not the case, and the variation in population distribution means that some regions which are small in area would have large populations and vice versa. Urban areas generally have higher population densities than rural areas. In the analysis of epidemiological data, it is common to specify an at risk population whose underlying spatial variation contributes to some of the spatial variation in the incidence of the disease be studied. For example, if the disease of interest was leukaemia in children aged 0-14 years, then the at-risk population used to normalise these disease counts would be the count of children aged 0-14.

Testing whether the values of the indicator are positive integers is straightforward:

```
test <- indicator >= 0 & ((indicator – floor(indicator)) == 0)
```

The resulting vector will have TRUE where the value of the indicator is a positive integer and FALSE otherwise. The observations where this test is FALSE are worthy of further scrutiny.

Ratio data: normalised

We start with ***aspatial testing***. If we assume that the distribution of values for a particular indicator followed a Normal distribution, then we could compute the first and second moments (the mean and variance). Approximately 95% of the data value could be expected to lie with ±1.96 standard deviations from the mean and approximately 99% of the data values would be within 2.58 standard deviations from the mean. Data value that were beyond ±3 standard deviations would be extremely unusual and worthy of examination. The problem we have with much socio-economic data is that, counterintuitively, it is rarely normally distributed. Usually there is a long tail to the right of the distribution of values. With income measurements, most people do not earn a great deal of money, and a few people earn a great deal – there are many poor, and just a few rich. Most houses are of moderate size, but a few are very large. The tests for outliers should take this into account and make few distributional assumptions. Tests such as those due to Dixon, Grubbs, and Cochran make the assumption that our data is drawn from a normal distribution, as do several others; for this reason we have not used them.

Distribution free estimates of the centre and spread of a distribution are represented by the median and interquartile range. The median is that value at which 50% of the observations have values below the median. The quartiles divide the distribution so that an equal number of observations have values between the three quartiles: if there were 100 observations in the data set, there would be 25 observations each in the ranges define by the quartiles and median.

A ***univariate*** filter is to consider the components of a boxplot, a technique developed by Tukey (1977). The interquartile range is the difference in values between the upper quartile and the lower quartile. In Tukey's nomenclature the quartiles are referred to as hinges, and the difference between them as the H-spread. The values of the fences are 1.5 times the H-spread beyond each hinge; these correspond roughly to the 99% confidence interval about the median. Any observations whose values are greater than that of upper fence and below that of the lower fence are held to be 'outside': that is, unusual. The advantage of the median based measures is that they are robust to distributional skew.

There are a number of ***multivariate*** alternatives. A simple method might be to postulate a linear relationship between a particular indicator and some or all of the other indicators in the dataset. In this case a linear model could be fitted and the residuals from this model tested. The residual is the difference between the observed value of some variable and the value predicted by the model. These can be standardised, such that residuals with a value greater than 2.58 (1% significance level) might be regarded as unusually high. This approach does require a model to be specified; this would require project domain specific understanding and the model may not be theoretically sound. A naive filter is to regress each variable of a set of variables on the other in turn, noting those which yield large standardised residuals. With m variables in a subset, we can tally for each region, the proportion of the m variables which yield large residuals when the are used as the response variable. A region where this proportion is close to 1 may be well be worthy of closer examination as being in some sense anomalous.

Many univariate tests consider the distance of an object from some central point; in the case of the tests which assume normality, this is the mean, and we are attempting to decide whether an object has a value with an unusually large *deviation* from the mean. A second approach to detecting outliers in high dimensional data is consider whether a a multivariate measure of distance is possible. If the variables in the analysis are uncorrelated, then the Euclidean distance from each object from the mean centre, could be computed. However, if the variables are correlated, then the space is not Euclidean, and some compensation is required. Mahalonobis' distances take the covariance structure of the data into account and may be used as measures of distance. Mahalonobis distance can be computed thus:

$$M_i = \sqrt{(\mathbf{x}_i - \overline{\mathbf{x}})^T V^{-1} (\mathbf{x}_i - \overline{\mathbf{x}})}$$

… where M_i is the Mahalobis distance for the ith object, the terms in the brackets are the deviances of the ith object from the mean vector and V is the covariance matrix for the data. We can then examine the boxplot statistics for these data. Where data are a mixture of quantitative and qualitative variables, a generalised version of the Mahalonobis distance exists (de Leon and Carrière, 2002) Fizlmoser et al (2003) have developed an adaptive iterative procedure for outlier detection based on robust estimation of Mahalonobis distances which can be used.

A third approach would be to combine the indicators in a ***Principal Components Analysis***. Principal components represent a transformation of the original data which attempts to capture much of the original data variance through a smaller number of variables. Values of the new variables can be computed from the component structure and are known as component scores. The boxplot statistics approach outlined earlier in this section would be useful in determining whether an observation represented an outlier or not. It would perhaps be reasonable to confine attention to the scores from the first component, as it represents that which accounts for the largest proportion of the original variance in the dataset indicators under test. Again, however, it would be unwise to pitch some or all of the indicators in a supplied dataset into such an analysis unless there were sound theoretical reasons for the chosen combinations.

Spatial approaches

We can also take into account that fact that the data which is undergoing a check is tied to the spatial arrangement of NUTS units for a specific time period. An observation which was made by the geographer Tobler (1970) is that "everything is related to everything else, but near things are more related than distant things". With spatial data it may not be unreasonable to expect that data values for adjacent NUTS regions would be similar to each other. If this were not the case, then an observation which is very different from its neighbours, might be a candidate to being regarded as an outlier.

Cressie (1984:33) observes that: "Analyzers of spatial data should be suspicious of observartions when they are unusual with respect to their neighbours". Cressie's exploratory analysis which follows this prescription uses as an example data on coal ash measurements from Robena Mine Property in Pennsylvania. The measurement locations are organised into ia regular grid which allows the use of tabular techniques such as Tukey's median polish. The NUTS regions do not form a regular lattice where reach cell is square as has four rooks-casse neighbours, but are irregularly spaced and also are of a wide range of shapes and sizes.

A simple ***spatial test*** is to compute local Moran's I statistics and tests these. The local Moran's I for a zone i is computed thus:

$$I_i = \frac{x_i - \bar{x}}{\left(\frac{\sum_{j,j\neq i} w_{ij}}{n-1} - \bar{x}^2 \right)} \sum_{j,j\neq i} w_{ij} (x_j - \bar{x})$$

… where x_i is the value of the indicator for the ith zone, w_{ij} is a matrix of weights (usually 1 for neighbour, 0 otherwise), and I_i is the local Moran's I. This statistic has the useful property that the sum of all the local I_is is the measure of the global spatial autocorrelation in the data under test. If the value of the local I is negative, this indicates that the value of the indicator for zone I is very different from those of its immediate neighbours, and this may well suggest the presence of an outlier.

A second ***spatial test*** is due to Hawkins (1980). The idea is similar to the Moran test, but the form of the test statistic is somewhat different:

$$H_i = \frac{n(x_i - \overline{x})^2}{(n+1)s^2}$$

In the formula n is the number of neighbours with x-bar is the arithmetic mean of the values in the neighbouring regions. The s^2 term is the vaergae variance for equvalently sized neighbourhoods over the sampling space. In our implementation the locally weighted summary statistics are based on a bi-square kernel with a fixed local sample size of 0.1 of the total number of regions in the dataset. The test statistic is distributed as a χ^2 with 1 degree of freedom, so that the 5% critical value is 3.84, and the 1% critical value is 6.64. In this test the values in all regions are considered to be suspects (Rossi et al, 2002). Krige and Magri (1982) show how Hawkins' test may be applied to gold mining data.

Classification based approaches

Another approach to the identification of univariate or multivariate outliers is provided by a classification based approaches. Classification methods, sometimes refrred to as cluster analysis methods, attempt to group the observations into a small number a clusters basedf on measurements of the similarity or dissimilarity between the observations. There are a wide number of methods and a wide number of distance measurements. A similarity measure increases in value, the more similar two objects are, and vice versa. Euclidean distances represent measure of dissimiliarity, as do Mahalanobis' distances.

A useful measure of similarity is provided by Gower's (1971) similarity coefficient. This is a general purpose measure which provides measures of dissimilarity for binary, nominal, ordinal and ratio data. The coefficient is defined:

$$S_{ij} = \frac{\sum_k w_{ijk} s_{ijk}}{\sum_k w_{ijk}}$$

… where Sij is the similarity between objects I and j, measured over k variables. Lowercase is the contribution of the kth variable, and w is 1 is the comparison on the kth variable is valid.

For binary (0/1, yes/no, presence/absence) data Gower's measure is defined as:

	Value of variable k			
Object i	1	1	0	0
Object j	1	0	1	0
s	1	0	0	0
W	1	1	1	0

If both objects have the value 1, then s is 1, 0 otherwise; if either object has the value 1, then w is 1, 0 otherwise. Gower extends this for nominal data, so that if both objects are in the same group then s is 1. For ordinal and ratio data the definition is s_{ijk} is:

$$s_{ijk} = 1 - \frac{\left| x_{ik} - x_{jk} \right|}{\Delta_k}$$

… where Δ_k is the range of the kth variable.

Torgo (2007, 2010) uses this dissimilarity measures as the basis for a classification based method which gives the 'outlierness' for a set of data. This ranking is based on the trajectory of an object during an agglomerative hierarchical clustering method. One of the outputs is the probability than an object is a multivariate outlier.

Which test?

We have outlined a number of approaches to detecting outliers. Some are specific to a particular type of data, others are general. Some are mechanical, others are statistical. We can deal with one attribute at a time or consider several simultaneously. Penny and Joliffe (2001) suggest that relying on a single technique is unwise, and that the analyst should consider the results from several tests. Where possible we follow this prescription.

Robust Methods

A number of the outlier detection methods are based on statistics which assume underlying normality of the data. As Riply (2004) points out, outliers can play havoc with such statistics. The mean is a case in point: as any data value approaches ∞, then the mean approaches ∞ as well. The median, on the other hand is little affected by the moving of a single value to ∞. Indeed, the median can tolerate up to 50% of the observations being gorssly in error, where for the mean this breakdown point is 0%. If the distribution is not normal, the mean can be a poor estimator of the central tendency of the distribution. In recent years there has been a development of robust methods, which are less influenced by outliers than standard statistical approaches. Counter-intuitively, this makes them useful tools for detecting outliers.

A simple robust extention of the mean is the **Winsorised mean**. In the computation of this statistic, the largest k+1 values are replaced by the kth value, and the smallest k+1 values are replaced by the kth smallest value. A similar statistic is the **trimmed mean**, in which the k largest and k smallest values are removed, and the divisor is n-2k. However, if the distributions are not symmetric, neither estimate will be an ubaissed estimate of the tryue mean.

An alternative is provide by the **median**, and a range of statistics based with their origins in Tukey (1974). The boxplot statistics have already been introduced – these define the salient features of the boxplot: the median, the quartiles (and the interquartile range) and the **whiskers** (who extend to the median±1.5*interquatile range). If the distribution is normal, the whiskers encompass all but the upper and lower 0.36% of the distribution. Observations whose values lie beyond those of the whiskers can be considered to be outliers.

It is tempting to remove the outlying observations, but, again, as Riply (2004) points out: a shapr decision to remove an observation is wasteful of data, identiffying outliers is difficult in highly structured data, and rejecting outliers means that estimates of the variance will be underestimated. Alternatives approaches are attractive in these situtations.

A **univariate** filter is provided by the **boxplot** and its associated statistics. For **bivariate** data, the **bagplot** provides an extension into two dimensions. The bagplot relies on the Tukey median (a two-dimensional version of the median) and the halfspace depth (a two dimensional equivalent of the interqartile range). The bag is the area around the median containg data with the highest depth, and the fence inflates the bag by a factor of 3. Observations beyond the fence are considered to be outliers (Hubert and van der Vekken, 2002). The median, bag boundary and fence boundary can be visualised on a scatterplot. For a series of variables, bagplots can replace the scatterplots in a scatterplot matrix display, although the aberrant cases are perhaps more easily caught by the computations required for for bagplot, and used in the weight of evidence estimation for the dataset under test.

> ** We have implemented the boxplot as part of the standard univariate suite of tests for ratio and count data

> ** We have implemented the bagplot as part of the bivariate suite of tests – outliers are extracted from the m(m-1)/2 possible comparisons of m ratio or count variables.

In the case of **multivariate** data, we require robust versions of **regression** and **principal components analysis**. A number of approaches to the estimation of **robust regression** have been proposes since the *least median of squares* appeared in the 1980s. In a robust estimate the outlying observations are not discarded, but are included in the analysis. The regression estimator includes a weighting function to ensure that the influence of the most extreme cases is minimised; the algorithm in R uses a technique known as the M-estimator, which is due to Huber (1981). As the robust regressions do not have standardised residuals in the manner of those that are available for Gaussian linear regression, potential outliers in the regression residuals can be identified by inspection of the boxplot statistics.

> ** We have implemented a robust regression as part of the bivariate suite of tests. Outliers are determined by the computation of boxplot statistics on the raw residuals

Robust principal components analysis provides an alternative to the standard principal components analysis (Hubert, Rousseeuw, and Vanden Branden, 2003). Recall that the eigenvectors are based on maximising the variance and decomposing the covariance matrix .Both the variance and covariance are infleunced by outliers. The effect is that the first components (which accounts for the highest variances) are attracted towards outlying observations, and may not capture the variation in the other observations (Hubert, Rousseeuw and Vanden Branden, 2003). The goal of the robust method is the minimise the effects of the outlying observations through downweighting, without removing any of the aberrant data cases.

Robust spatial versions of the univariate, bivariate and multivariate statistics still represent something of a research frontier. Liu et al (2001) use the Jacknife to create an interpolation residual based index for outliers; however, their methods is coupled to the ArcINFO GIS software. They also note that for smal datasets the robust estimates of the summary statistics can only be estimated with sufficient observations.

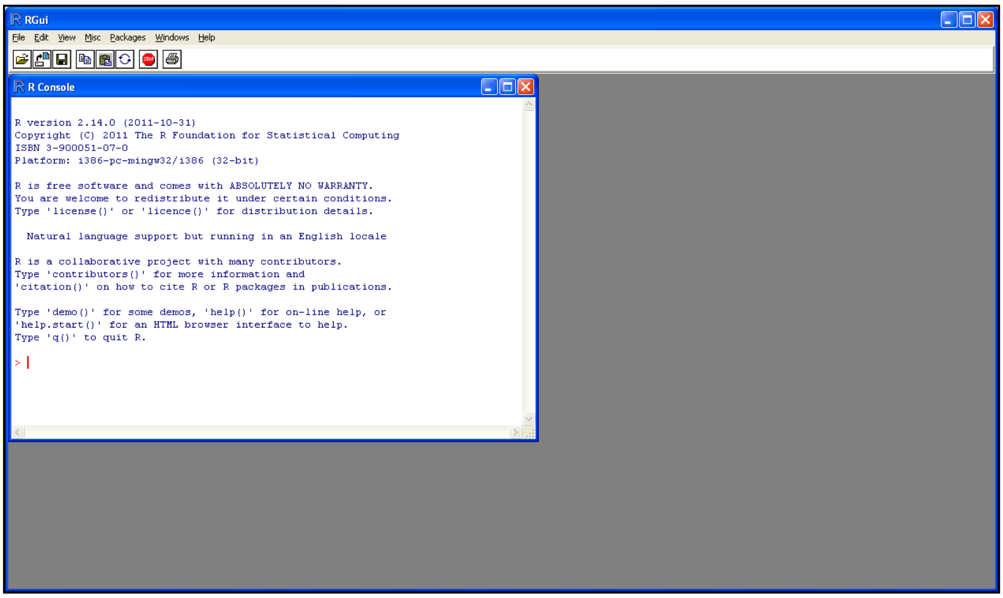
Robust versions of kriging-based geostatistical methods also exist. However, geostatistical methods largely involve the prediction of the values of an indicator at locations where samples have not been taken. For example, samples of air quality may have been taken at a variety of locations around a city, and geostatistical methods used to estimate the levels of pollutants at the mesh points of a regular lattice. This involves the estimation of a variogram which gives a picture of the spatial covariance structure of the data. The spatial weights for the estimation are drawn from a theoretical model fitted to the empirical variogram. Some of the procedure can be automated, but for the best estimates, a degree of interaction is desirable. Diagnostics for geostatistical models potentially involve interaction and are perhaps less suited to an automated environment (Bradley and Haslett, 1992). At the moment, the degree of expertise required to use geostatistical methods renders it a specialist subject for the ESPON database. However, this does not preclude their use in the outlier detection process.

Towards an implementation of the outlier check

The previous sections of the Technical report have outlined various methods for checking data for the presence or otherwise of anomalous values. In the M4D data flow model the outlier check takes place at the NCG following syntaxic and semantic checks on the metadata and the data.

In designing an implementation of the outlier check process, the desire has been (a) to keep interference with the data in the spreadsheet to a minimum and (b) the keep the amount of coding for each dataset to a minimum. Outlier checking is not a completely automated process, and a certain amount of interaction will be needed to draw conclusions from the check. We have already observed that outliers may arise from incorrect entry of the data or mistakes in computation; outleirs may also arise because some NITS regions are generalised very different from their neighbours or smaller units. This may require interaction with the supplier if we uncover data which might be anomalous. The stark and potentially accusatory nature of a printout of the R code output will require some context, and some interpretation. To be an outlier is a matter of degree, and that importance degree of outlyingness is something for the M4D team and the data supplier to agree on.

We have decided to implement the checking functions in the R system (R DCT, 2005). R is an open source environment with a rich body of functions for data manipulation and statistical analysis. In recent years, these have been added to with several packages for handling and analysing spatial data. Unlike SPSS the analyst primarily communicates with R through a *command line interface*: the R console. This demands a familiarity with the R language itself, and the syntax, structure and interrelationships between the various functions which are available.



```
R version 2.14.0 (2011-10-31)
Copyright (C) 2011 The R Foundation for Statistical Computing
ISBN 3-900051-07-0
Platform: i386-pc-mingw32/i386 (32-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> 
```

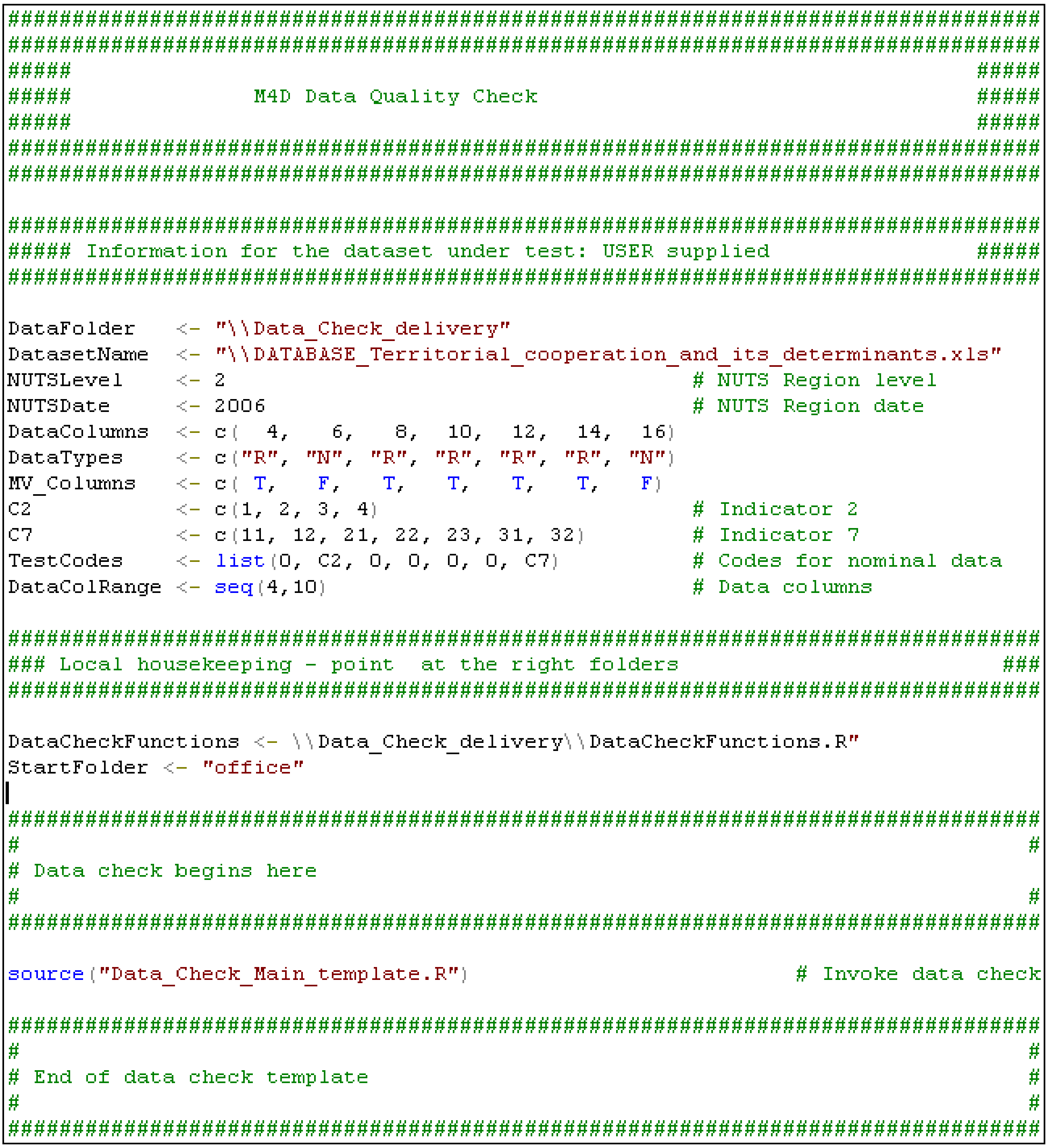
R will not do anything until instructed, and the > prompt will remain until the user enters an R instruction. As an example, to read a comma-separated-variable file into a data frame x, the user might enter:

X <- read.csv(“social_class_nuts3.csv”)

… the result of which will be to transfer the contents of the file *social_class_nuts3.csv* into the data frame *x*. The rows in the data frame will correspond to the observations (NUTS3 regions in this case) and the columns to the indicators or variables (measures of social class).

Clearly entering the individual instructions to copy the data from the Excel spreadsheets which will be supplied as the data interchange medium and undertake the analysis would be an onerous task, so R permits the creation of functions which permit various related tasks to be group together as effectively a new instruction written in the R language. The new functions can be grouped together in text files. To invoke the instructions in a file the **source()** command is used, as in the following example.

To achieve the goals for the outlier check we have written a hierarchical set of checking functions which a supplied with appropriate control information. We have experimented with extracting this control information from the metadata in the spreadsheets, but the most reliable approach has been to code it by hand.



```
#################################################################################
#################################################################################
#####                                                                       #####
#####              M4D Data Quality Check                                   #####
#####                                                                       #####
#################################################################################
#################################################################################

#################################################################################
##### Information for the dataset under test: USER supplied                 #####
#################################################################################

DataFolder   <- "\\Data_Check_delivery"
DatasetName  <- "\\DATABASE_Territorial_cooperation_and_its_determinants.xls"
NUTSLevel    <- 2                                     # NUTS Region level
NUTSDate     <- 2006                                  # NUTS Region date
DataColumns  <- c(  4,   6,   8,  10,  12,  14,  16)
DataTypes    <- c("R", "N", "R", "R", "R", "R", "N")
MV_Columns   <- c( T,   F,   T,   T,   T,   T,   F)
C2           <- c(1, 2, 3, 4)                         # Indicator 2
C7           <- c(11, 12, 21, 22, 23, 31, 32)         # Indicator 7
TestCodes    <- list(0, C2, 0, 0, 0, 0, C7)           # Codes for nominal data
DataColRange <- seq(4,10)                             # Data columns

#################################################################################
### Local housekeeping - point  at the right folders                          ###
#################################################################################

DataCheckFunctions <- \\Data_Check_delivery\\DataCheckFunctions.R"
StartFolder <- "office"

#################################################################################
#                                                                               #
# Data check begins here
#                                                                               #
#################################################################################

source("Data_Check_Main_template.R")                         # Invoke data check

#################################################################################
#                                                                               #
# End of data check template                                                    #
#                                                                               #
#################################################################################
```

Each spreadsheet that is supplied is composed of four separate worksheets

Dataset	description of the dataset, project and supplier
Indicator	metadata concerning the individual indicators
Source	information on the source of the data
Data	the data itself: rows represent the observations; Columns represent the indicators. Each indicator appears as a column pair, the first containing the data values and the second a code linking to the Source metadata

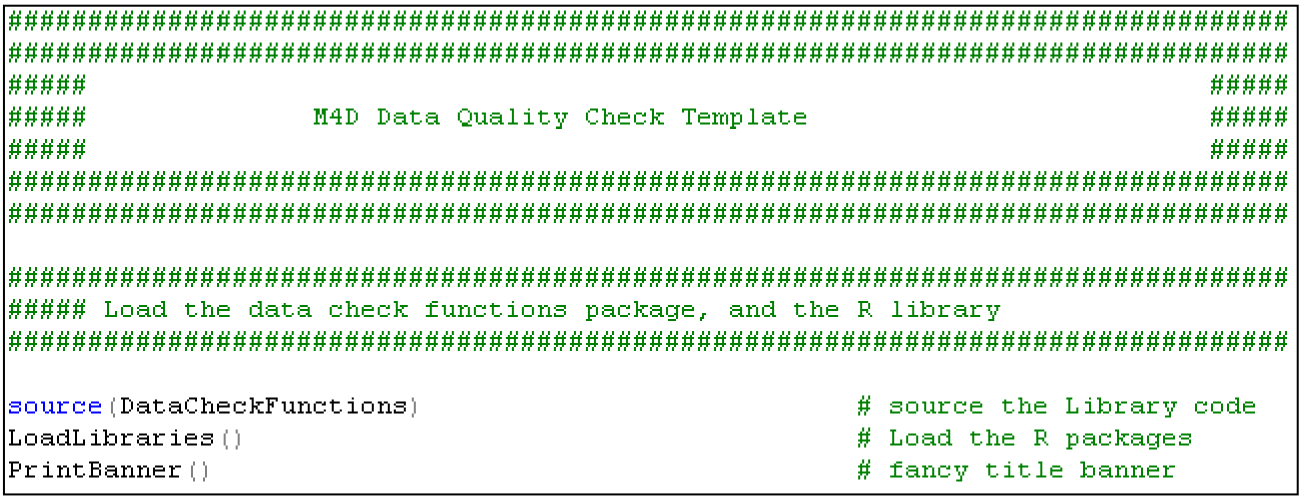
The spreadsheet is read directly by R and the contents of each worksheet copied into a separate dataframe. No manual intervention is required. The control information is obtained from a prior manual inspection of the spreadsheet. There are 9 main piece of control information.

Variable	Content
DataFolder	Points at the location of the folder which contains the spreadsheet to be checked.
DatasetName	The name of the Excel spreadsheet in DataFolder
NUTSLevel	The level of the NUTS regions. Currently {0 | 1 |2 | 3} but this will be expanded as different areal units are encountered (UMZ, Corine regions &c)
NUTSDate	Date of the spatial units – there are different shapefiles for the different NUTS epochs
DataColumns	A vector of numeric indices which point to the columns which contain data for checking
DataTypes	A vector of data type indicators (R: ratio, N: Nominal in this case). This vector has the same number of elements as DataColumns
MVColumns	A vector indicating which variables will be used in the multivariate checking (T: yes, F: no). Again, this vector is the same length as the DataColumns Vector
C*n*	A vector of the allowable codes for the *n*th variable if this is nominal data. In the example column 2 and 7 are typologies with 4 codes for the first and 7 codes for t he second)
TestCodes	This is an R list used to transfer the typology codes for the nominal variables. This list is the same length as the DataColumns vector.

The 'Local Housekeeping' section supplies pointers to the locations in Dropbox where we are working. Three researchers in the NCG have been collaborating on the development of the data check process, and we have use Dropbox to coordinate our activities. The location of the Dropbox folder depends on the version of Windows being used, and the username of the researcher. This information is used to extract the spreadsheet, NUTS indices and any geometry files. The rest of the R code and functions are in two linked files – the first (Data_Check_Main_tempate.R) is needed to begin the data check procedure.

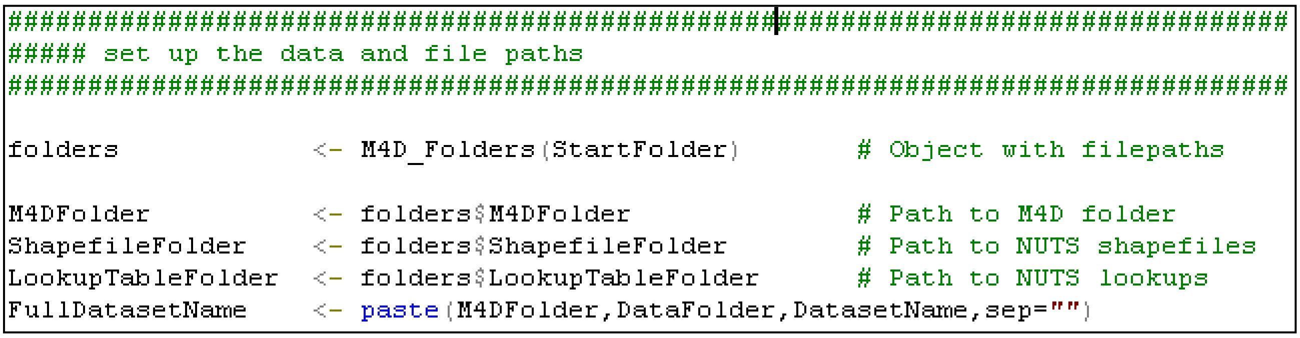
Data Check Function

The main driver code is contained in the file Data_Check_Main_Template.R. It initialises the check, ensures that the correct geometries are loaded, carries out some initial consistency checks, and invokes the various checking strategies, passing ht control information where necessary.



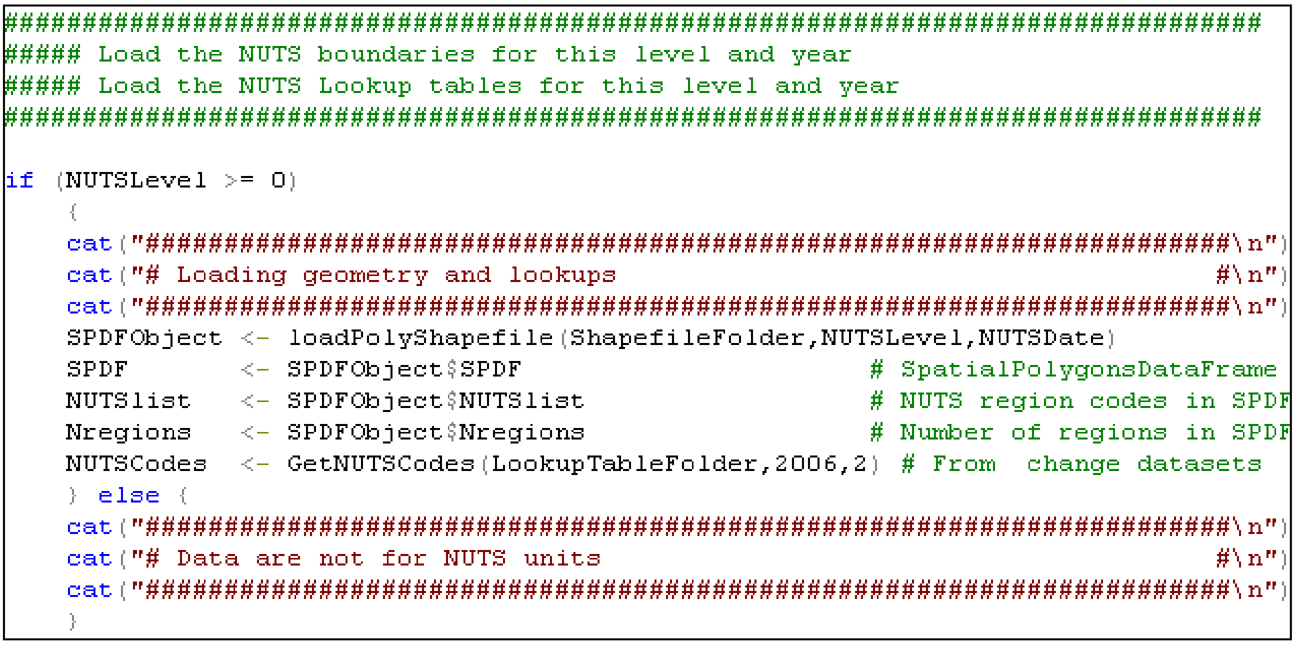
```
################################################################################
################################################################################
#####                                                                     #####
#####              M4D Data Quality Check Template                        #####
#####                                                                     #####
################################################################################
################################################################################

################################################################################
##### Load the data check functions package, and the R library
################################################################################

source(DataCheckFunctions)                           # source the Library code
LoadLibraries()                                      # Load the R packages
PrintBanner()                                        # fancy title banner
```

The first section loads the NCG function library into R (in the file DataCheckFunctions.R), loads the package libraries used in these functions, and then prints a banner to indicate the start of the checking process. The banner also prints the date and time that the check was run, if this is needed for subsequent audit purposes.



```
################################################################################
##### set up the data and file paths
################################################################################

folders            <- M4D_Folders(StartFolder)       # Object with filepaths

M4DFolder          <- folders$M4DFolder              # Path to M4D folder
ShapefileFolder    <- folders$ShapefileFolder        # Path to NUTS shapefiles
LookupTableFolder  <- folders$LookupTableFolder      # Path to NUTS lookups
FullDatasetName    <- paste(M4DFolder,DataFolder,DatasetName,sep="")
```

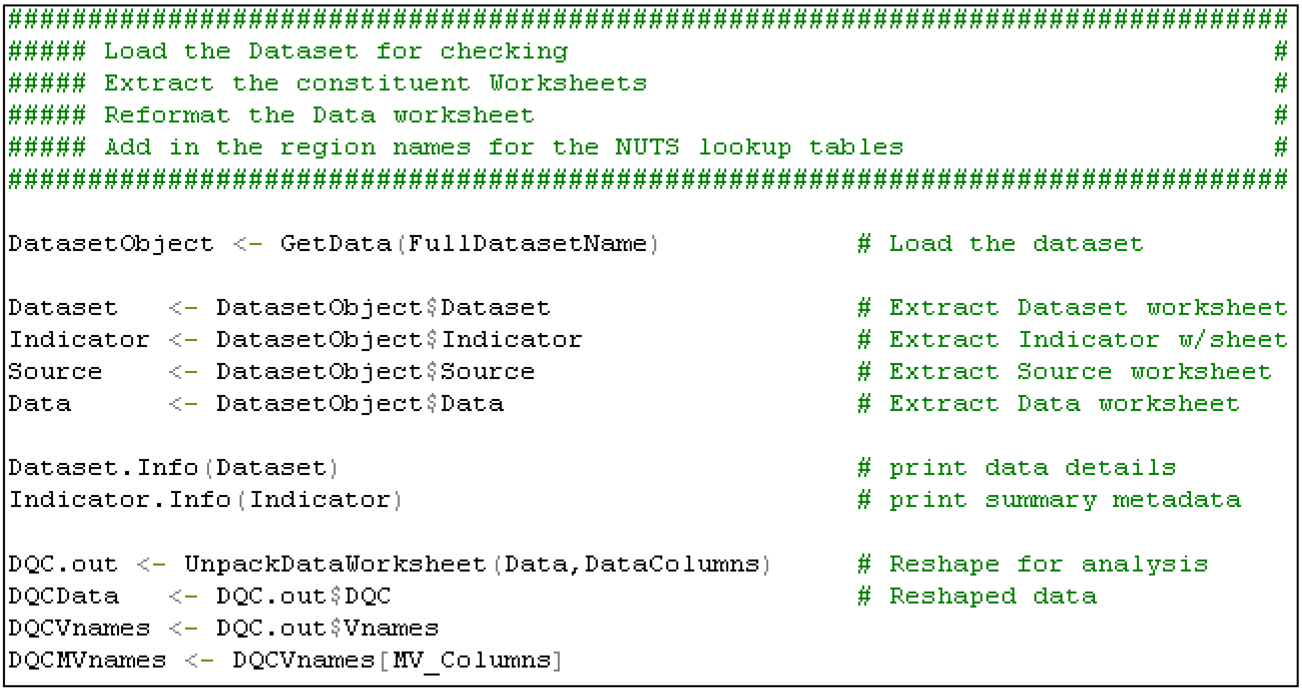
The location of the StartFolder is used by the M4D_Folders to return a lsit of points to the locations of the folders for (a) the data for checking, (b) the NUTS shapefiles and (c) the NUTS lookup tables. The full dataset name included not just its name, but the complete path to its location. This is the file that will be read into R.



```
################################################################################
##### Load the NUTS boundaries for this level and year
##### Load the NUTS Lookup tables for this level and year
################################################################################

if (NUTSLevel >= 0)
    {
    cat("######################################################################\n")
    cat("# Loading geometry and lookups                                       #\n")
    cat("######################################################################\n")
    SPDFObject <- loadPolyShapefile(ShapefileFolder,NUTSLevel,NUTSDate)
    SPDF       <- SPDFObject$SPDF                        # SpatialPolygonsDataFrame
    NUTSlist   <- SPDFObject$NUTSlist                    # NUTS region codes in SPDF
    Nregions   <- SPDFObject$Nregions                    # Number of regions in SPDF
    NUTSCodes  <- GetNUTSCodes(LookupTableFolder,2006,2) # From  change datasets
    } else {
    cat("######################################################################\n")
    cat("# Data are not for NUTS units                                        #\n")
    cat("######################################################################\n")
    }
```

If the control information indicates the level and date for the NUTS units. This section of the code is responsible for copying the relevant geometry shapefile into the internal data structure used for spatial data in R: this structure is known as a Spatial Polygons data Frame, or SPDF. The SPDF contains not only the geometry data which describes the boundaries of the NUTS polygons but also some attribute data including the NUTS code, the polygon area, the NUTS level and date, and the name of the NUTS region. Some of the names are in locale specific script – not all this translate into the internal structure successfully, and we are investigating why this is so and what needs be undertaken as a workaround.

The loadPolyShapefile function returns not only the SPDF object, but also the list of NUTS codes which are present, and counts the number of NUTS regions for the combination of level and date. A second function retrieves a corresponding list of NUTS codes and name for the level/date.

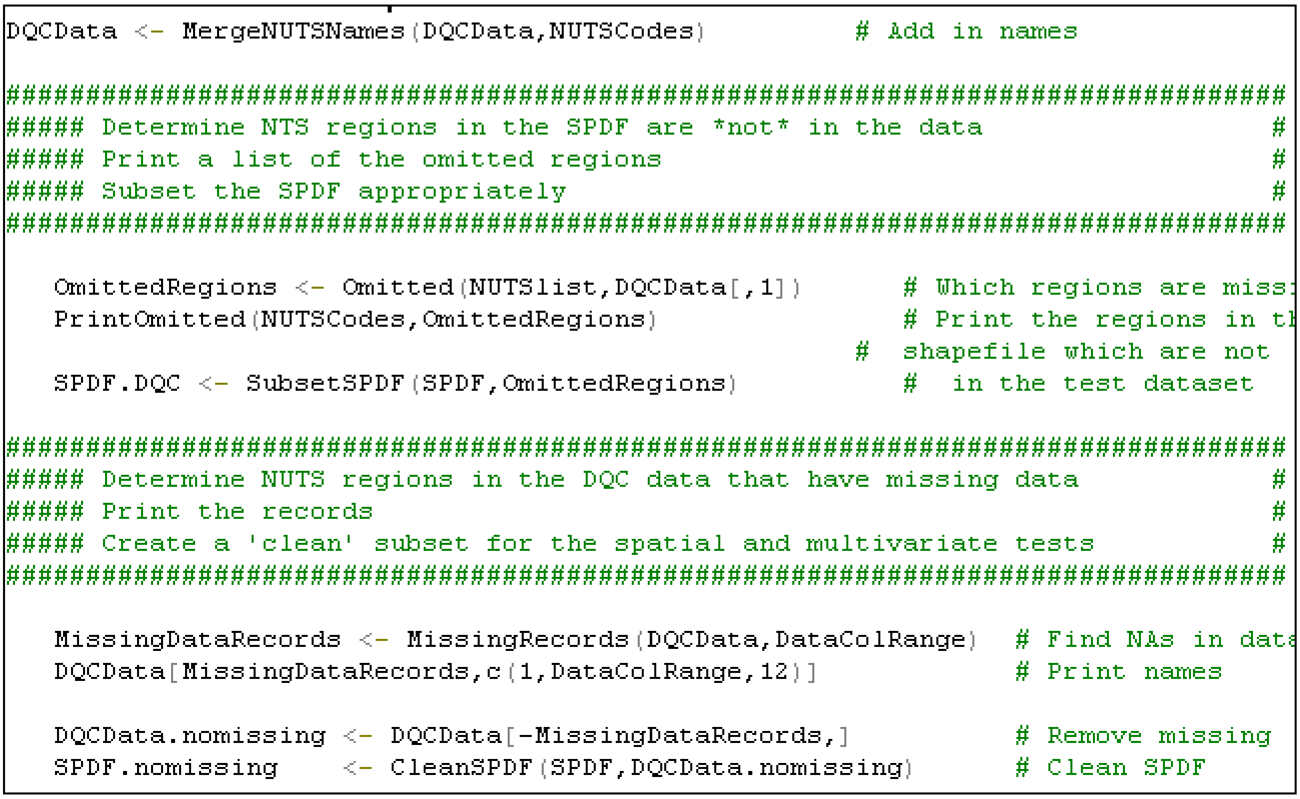


```
################################################################################
##### Load the Dataset for checking                                            #
##### Extract the constituent Worksheets                                       #
##### Reformat the Data worksheet                                              #
##### Add in the region names for the NUTS lookup tables                       #
################################################################################

DatasetObject <- GetData(FullDatasetName)            # Load the dataset

Dataset   <- DatasetObject$Dataset                   # Extract Dataset worksheet
Indicator <- DatasetObject$Indicator                 # Extract Indicator w/sheet
Source    <- DatasetObject$Source                    # Extract Source worksheet
Data      <- DatasetObject$Data                      # Extract Data worksheet

Dataset.Info(Dataset)                                # print data details
Indicator.Info(Indicator)                            # print summary metadata

DQC.out <- UnpackDataWorksheet(Data,DataColumns)     # Reshape for analysis
DQCData   <- DQC.out$DQC                             # Reshaped data
DQCVnames <- DQC.out$Vnames
DQCMVnames <- DQCVnames[MV_Columns]
```

The GetData function invokes some low level ODBC functions which open the Excel spreadsheet, copy the worksheets into R data frames, and closes the spreadsheet. No manipulation beyond the sqlFetch instruction has taken place, so some data manipulation is required to reshape the dataframe for the data checking functions.

Some summary information to identify the dataset and the indicators it contains is extracted from the metadata. The reshaping consist of extracting the NUTS codes, date and level, and the columns which contain the data and copying them into a new data frame. In the spreadsheet the first row is used for the name of the indicator column. However, for some data this is not unique. The second and third rows may contain one or more dates, so these are concatenate with the corresponding indicator name to give a unique column name. The first three rows are omitted from the data frame, and the column names in data frame are then taken from the unique names which have been created in this step.

The indicator names are written to a vector – this will be sued in the subsequent sections for the data checking, and the corresponding DataType and TestCodes lists used to select and operate the appropriate checking procedures. The subset of columns to be used for the multivariate data checked is also created in this step.

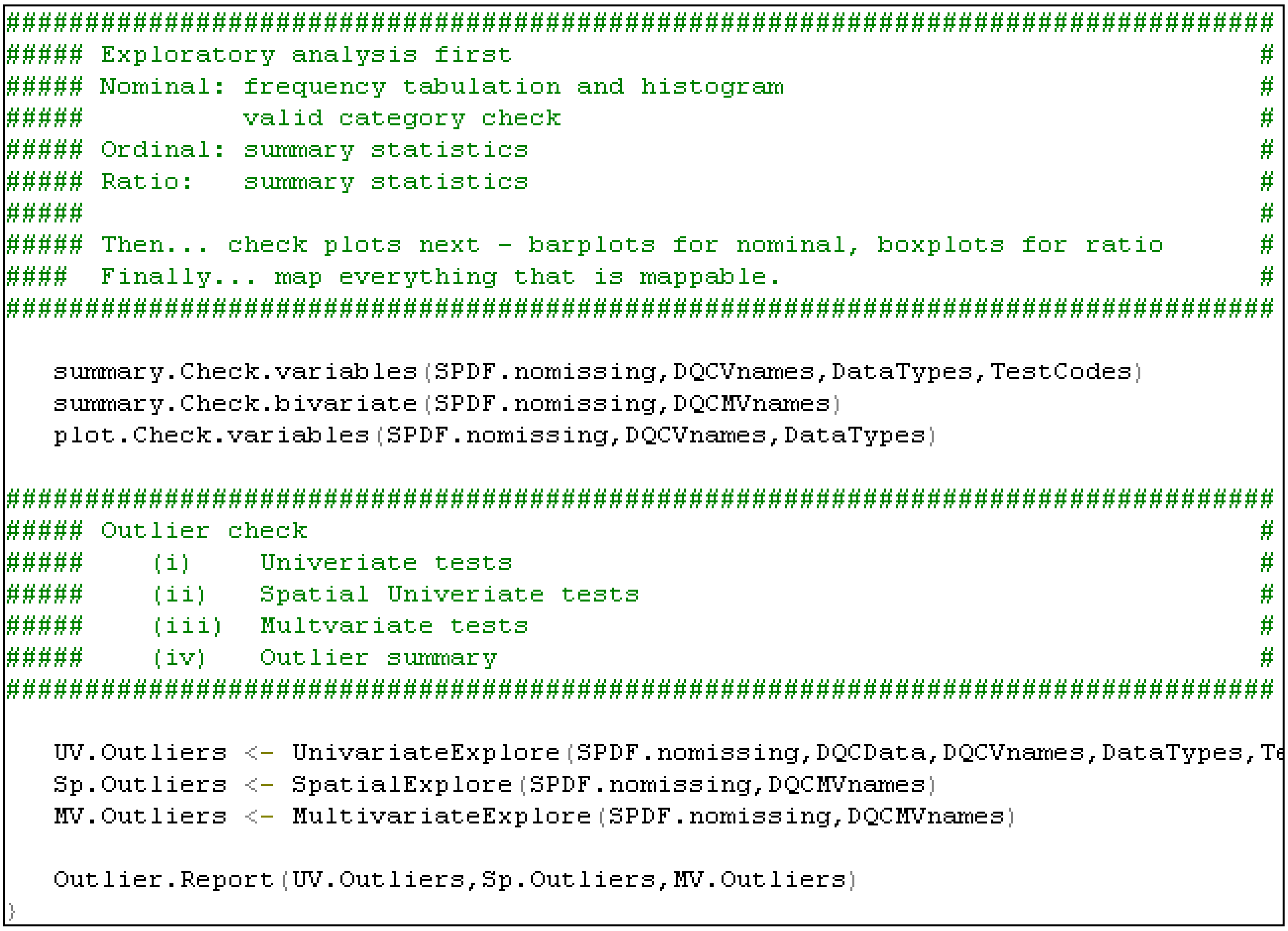


```
DQCData <- MergeNUTSNames(DQCData,NUTSCodes)         # Add in names

################################################################################
##### Determine NTS regions in the SPDF are *not* in the data                  #
##### Print a list of the omitted regions                                      #
##### Subset the SPDF appropriately                                            #
################################################################################

   OmittedRegions <- Omitted(NUTSlist,DQCData[,1])      # Which regions are missi
   PrintOmitted(NUTSCodes,OmittedRegions)               # Print the regions in th
                                                     #  shapefile which are not
   SPDF.DQC <- SubsetSPDF(SPDF,OmittedRegions)          #  in the test dataset

################################################################################
##### Determine NUTS regions in the DQC data that have missing data            #
##### Print the records                                                        #
##### Create a 'clean' subset for the spatial and multivariate tests           #
################################################################################

   MissingDataRecords <- MissingRecords(DQCData,DataColRange)  # Find NAs in data
   DQCData[MissingDataRecords,c(1,DataColRange,12)]            # Print names

   DQCData.nomissing <- DQCData[-MissingDataRecords,]          # Remove missing
   SPDF.nomissing    <- CleanSPDF(SPDF,DQCData.nomissing)      # Clean SPDF
```

Not all projects use the complete set of NUTS regions for a given level/date. Those which are not present are identified and removed from the Spatial Polygons Data Frame prior to the final data checking step.

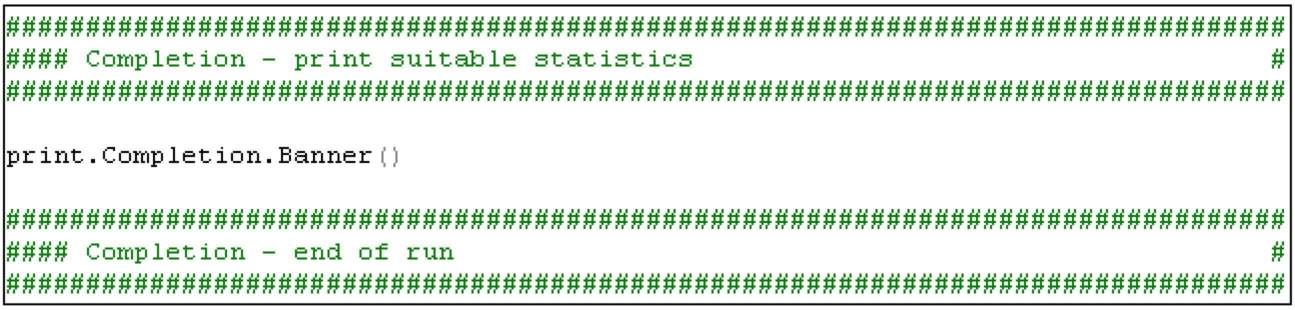
It sometimes happens in data collation that a value for a particular region is not available at the time of the analysis or for the duration of the project. The NA combination is often used to signal a missing value, although in some cases a cell is left unfilled. R creates an implicit NA for the empty cells. It is conventional to organise the omission of such data in computations. For univariate analysis this can be done on an indicator by indicator basis, although it does mean that potentially different samples are used for different indicators. With multivariate analysis, an record with missing data in any candidate cell is omitted from the analysis: this is known as casewise deletion.

The final step is the casewise removal of any records that have missing data across the complete range of indicators. We have encountered one or two cases in the data we have seen. The removal is both from the 'aspatial' data frame and the spatial data frame.



```
################################################################################
##### Exploratory analysis first                                               #
##### Nominal: frequency tabulation and histogram                              #
#####          valid category check                                            #
##### Ordinal: summary statistics                                              #
##### Ratio:   summary statistics                                              #
#####                                                                          #
##### Then... check plots next - barplots for nominal, boxplots for ratio      #
####  Finally... map everything that is mappable.                              #
################################################################################

   summary.Check.variables(SPDF.nomissing,DQCVnames,DataTypes,TestCodes)
   summary.Check.bivariate(SPDF.nomissing,DQCMVnames)
   plot.Check.variables(SPDF.nomissing,DQCVnames,DataTypes)

################################################################################
##### Outlier check                                                            #
#####    (i)     Univeriate tests                                              #
#####    (ii)    Spatial Univeriate tests                                      #
#####    (iii)   Multvariate tests                                             #
#####    (iv)    Outlier summary                                               #
################################################################################

   UV.Outliers <- UnivariateExplore(SPDF.nomissing,DQCData,DQCVnames,DataTypes,Te
   Sp.Outliers <- SpatialExplore(SPDF.nomissing,DQCMVnames)
   MV.Outliers <- MultivariateExplore(SPDF.nomissing,DQCMVnames)

   Outlier.Report(UV.Outliers,Sp.Outliers,MV.Outliers)
}
```

The data check proper consists of two parts: an exploratory analysis of the indicators in the data and an a confirmatory analysis to identify any potential outliers. The exploratory analysis includes the provision of appropriate summary statistics and tabulations, bivariate correlations, boxplots of the ration/count data, barplots of the typology data, and maps of all the indicators. The confirmatory analysis applies a battery of tests: univariate, spatial, and multivariate. The final action is to print a summary of the results in the creation of the weight of evidence for a region being outlying in some sense. .



```
################################################################################
#### Completion - print suitable statistics                                    #
################################################################################

print.Completion.Banner()

################################################################################
#### Completion - end of run                                                   #
################################################################################
```

The final action is to write the summaries to a file, and print a completion banner. The banner signals that the data check run have completed successfully and prints the time that this occurred. This is potentially useful for audit purposes; however, the R functions will occasionally terminate prematurely if an unexpected data characteristic has occurred.

Data check function library

The library of data check functions includes the intermediate level functions driving the univariate, spatial and multivariate checks, as well a low level function to invoke particular tests. At the time of writing this consists of some 900 lines of R instructions and associated comments.

While is it based on the R code created by Paul Harris for the Phase I project, the library has been created from scratch in accordance with the redesign of the outlier check in mind. Further functions will be added to enlarge the range of possible tests.

Documentation of this library is in progress. Copies can be made available through the ESPON Coordination Unit.

Example data check session

An example session is shown using some example data from an ESPON Project. The dataset was supplied by the Coordination Unit and contains examples of typology and ratio date for NUTS3 regions in their 2006 version.

Within the dataset there are 7 indicators: two a typology variables; 4 are principal component scores, and one consists of unstandardised residuals from a regression model. Current ideas about M4D are that regression residuals and principal component scores are perhaps unhelpful as indicators; the team would rather the constituent variables were stored, allowing an analyst to replicate the analysis undertaken by the project.

The control information shown in the previous section was derived from this dataset. There are some observations we would make concerning the data. First, regression residuals should be homoskedastic: they should be random, have a mean of zero, and be independently and identically distributed; furthermore any random sample should have the same variance. As they apply to spatial units, they should not exhibit any spatial autocorrelation. We have not checked for any misspecification in the model structure as revealed in the residuals, although the team would be able to offer statistical advice to any project using regression methods.

Second, four of the variables are principal component scores. We know little about the variables which were subject to the principal components analysis. We would assume that the components shown have eigenvalues which are greater than 1; as we examine the correlation structure of the multivariate data we can check for any departure from orthogonality. However, the database team would rather have the original variables – this allows any other analyst to replicate the PCA, or use a different form of PCA.

Comments on the output will appear in *italic*.

The first section displays the banner: this shows the time and date that the quality check was made.

```
+================================================================+
###                                                           ###
### M4D: Data Quality Check                                   ###
###                                                           ###
### National Centre for Geocomputation                        ###
### National University of Ireland Maynooth                   ###
### Maynooth, Co Kildare, Ireland                             ###
###                                                           ###
+================================================================+
Run on:  Thu Jun 28 17:19:30 2012
```

The next section presents the information from, first, the ***Dataset*** *worksheet, and second, some of the metadata from the* ***Indicator*** *worksheet. It would be a desirable development to be able to pick the code lists for the typology variables from the metadata directly without having to include them in the control information. This should be a relatively simple task.*

```
+================================================================+
### Dataset identification                                    ###
+================================================================+
Dataset name:         Territorial cooperation and its determinants
Project:              XXXXX
Abstract:            Typology type of data related to territorial cooperation
Unique Resource Identifier  XXXXX_terr_coop_det_2012_v1
```

```
+===============================================================+
### Summary information from the metadata                     ###
+===============================================================+
Indicator Name          Data Type     Data Values
RES_DIS_RAN             float         km
DET_TYP                 integer       types 1-4
PCA_1_CORE              float         none
PCA_2_ATTRACT           float         none
PCA_3_PROBLEM           float         none
PCA_4_METRO             float         none
REG_TYP                 integer       types 11, 12, 21, 22, 23, 31, 32
```

We next check the list of NUTS units in the Data worksheet against the list of NUTS units extracted from the lookup list. A number of studies take a sample of NUTS units, others omit some of the detached regions. In this example, Ceuta and Melilla are omitted, as are the French territories of Guadeloupe, Martinique, Guyane and Réunion. The Azores likewise were not included in the study. These are not errors, but they are pertinent information for this particular subset of the EU regions.

```
MISSING REGIONS in the Data for Checking
Regions in the NUTS list not in the data for checking
     Code                       Name Level
781  ES63   Ciudad Autónoma de Ceuta     2
783  ES64 Ciudad Autónoma de Melilla     2
956  FR91                 Guadeloupe     2
958  FR92                 Martinique     2
960  FR93                     Guyane     2
962  FR94                    Réunion     2
1456 PT20 Região Autónoma dos Açores     2
```

The regions present in the spreadsheet are also checked against the list of regions in the shapefile. There are extra regions in the shapefile. Macedonia is an example it submitted its application for EU membership in 2004, and has been a candidate for accession since 2005. The process of accession has yet to be completed, so as it had accession status, the boundaries are present in the shapefile. Note that this seven units noted above are also included in the list of omits. The other omitted countries are Croata (HR), Liechtenstein (LI) and Turkey (TR). These regions will be removed from the shapefile representation in R.

```
Regions in the Shapefile not in the data for checking
ES63 ES64 FR91 FR92 FR93 FR94 HR01 HR02 HR03 LI00 MK00 PT20 TR10
TR21 TR22 TR31 TR32 TR33 TR41 TR42 TR51 TR52 TR61 TR62 TR63 TR71
TR72 TR81 TR82 TR83 TR90 TRA1 TRA2 TRB1 TRB2 TRC1 TRC2 TRC3
```

The first section of the univariate exploratory analysis is driven by the datatypes. For the ratio data the check returns a five number summary. These mark summary measures of the distribution. We can see that the regression residuals (RES_DIS_RAN) are not standardized, and may be skew (the mean is some distance from the median). The principal components are distributed around a mean of about zero, and are not too skew. The last variable is a typology with seven categories. There are appears to be a single instance of 0 has been entered as a category value. This is not an obviously correctable error, and we would check this with the data supplier.

```
+===============================================================+
### Exporatory data summaries                                  ###
+===============================================================+
Variable:        RES_DIS_RAN
Data Type:       R
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-957.800 -256.500  -35.750    3.704  196.600 2445.000
-----------------------------------------------------------------
Variable:        DET_TYP
Data Type:       N
  1   2   3   4
 54  49 105  69
Categories found are:  1 2 3 4
-----------------------------------------------------------------
Variable:        PCA_1_CORE
Data Type:       R
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-2.0830 -0.5734  0.3008  0.1183  0.8610  2.3000
-----------------------------------------------------------------
Variable:        PCA_2_ATTRACT
Data Type:       R
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-1.92100 -0.68040 -0.04285  0.04166  0.61350  4.73600
-----------------------------------------------------------------
Variable:        PCA_3_PROBLEM
Data Type:       R
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-3.32800 -0.53000 -0.05204 -0.02480  0.52280  3.32100
-----------------------------------------------------------------
Variable:        PCA_4_METRO
Data Type:       R
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-1.24800 -0.54670 -0.24530 -0.01657  0.18830  6.52000
-----------------------------------------------------------------
Variable:        REG_TYP
Data Type:       N
 0 11 12 21 22 23 31 32
 1 31 21 26 39 16 54 89
Categories found are:  0 11 12 21 22 23 31 32
-----------------------------------------------------------------
```

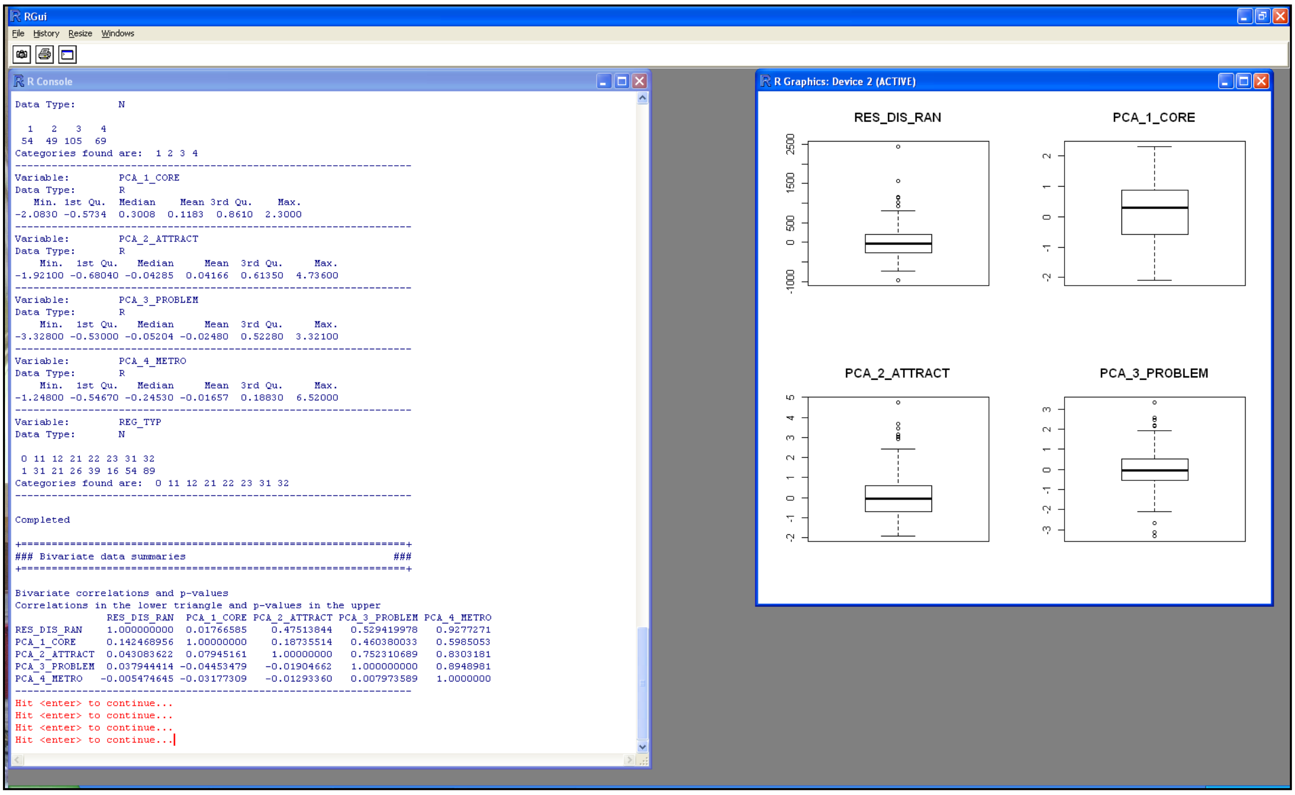
*In the next section are some bivariate measures – Pearson product moment correlation coefficients. These need some interpretation given our knowledge of the types of data that are present in the file. The table below presents the correlations between every pair of variables in the **multivariate list**. The correlations are in the lower triangle of the matrix; the leading diagonal is blank [the correlations would be 1], and the upper triangle presents the p values for the hypothesis that the coefficient is not significantly different from zero.*

As expected the principal component appear to be orthogonal – the correlations are close to zero, and the p values are all greater than 0.05. The regression residuals are also uncorrelated with the principal components – this might be entirely by chance, or it may be that the y variable is included in the first component.

```
+================================================================+
### Bivariate data summaries                                   ###
+================================================================+

Bivariate correlations and p-values
Correlations in the lower triangle and p-values in the upper triangle
              RES_DIS_RAN  PCA_1_CORE PCA_2_ATTRACT PCA_3_PROBLEM PCA_4_METRO
RES_DIS_RAN                0.01766585    0.47513844   0.529419978   0.9277271
PCA_1_CORE    0.142468956                0.18735514   0.460380033   0.5985053
PCA_2_ATTRACT 0.043083622  0.07945161                 0.752310689   0.8303181
PCA_3_PROBLEM 0.037944414 -0.04453479   -0.01904662                 0.8948981
```

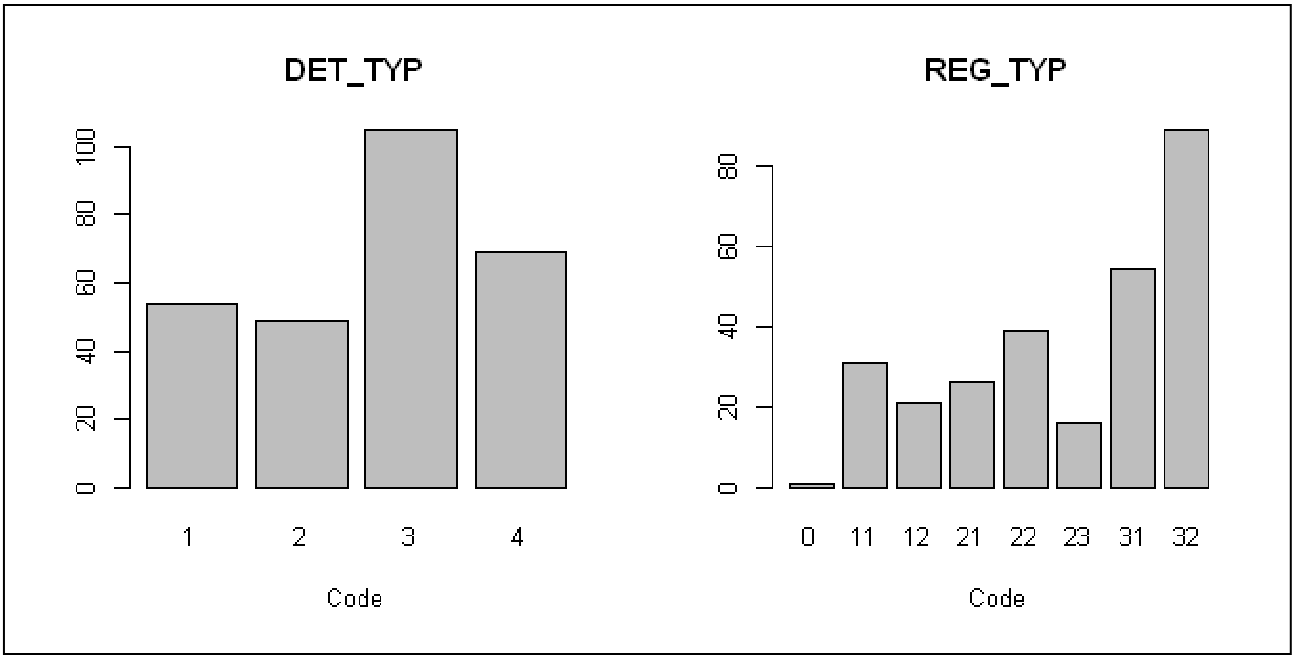
```
PCA_4_METRO   -0.005474645 -0.03177309    -0.01293360    0.007973589
```

The exploratory analysis is accompanied by appropriate visualisations. For the ratio variables we use boxplots, and for the typology variables we use barplots. The principal components behave as expected, although those later ones that account for less of the variance show a tendency to outlying observations. The regression residuals appear to have a small of observations in the right tail which will be identified as outliers. If they are not outliers on several of the other tests, then they will have little evidence in favour of their being consistently unusual. The display below shows the R GUI.

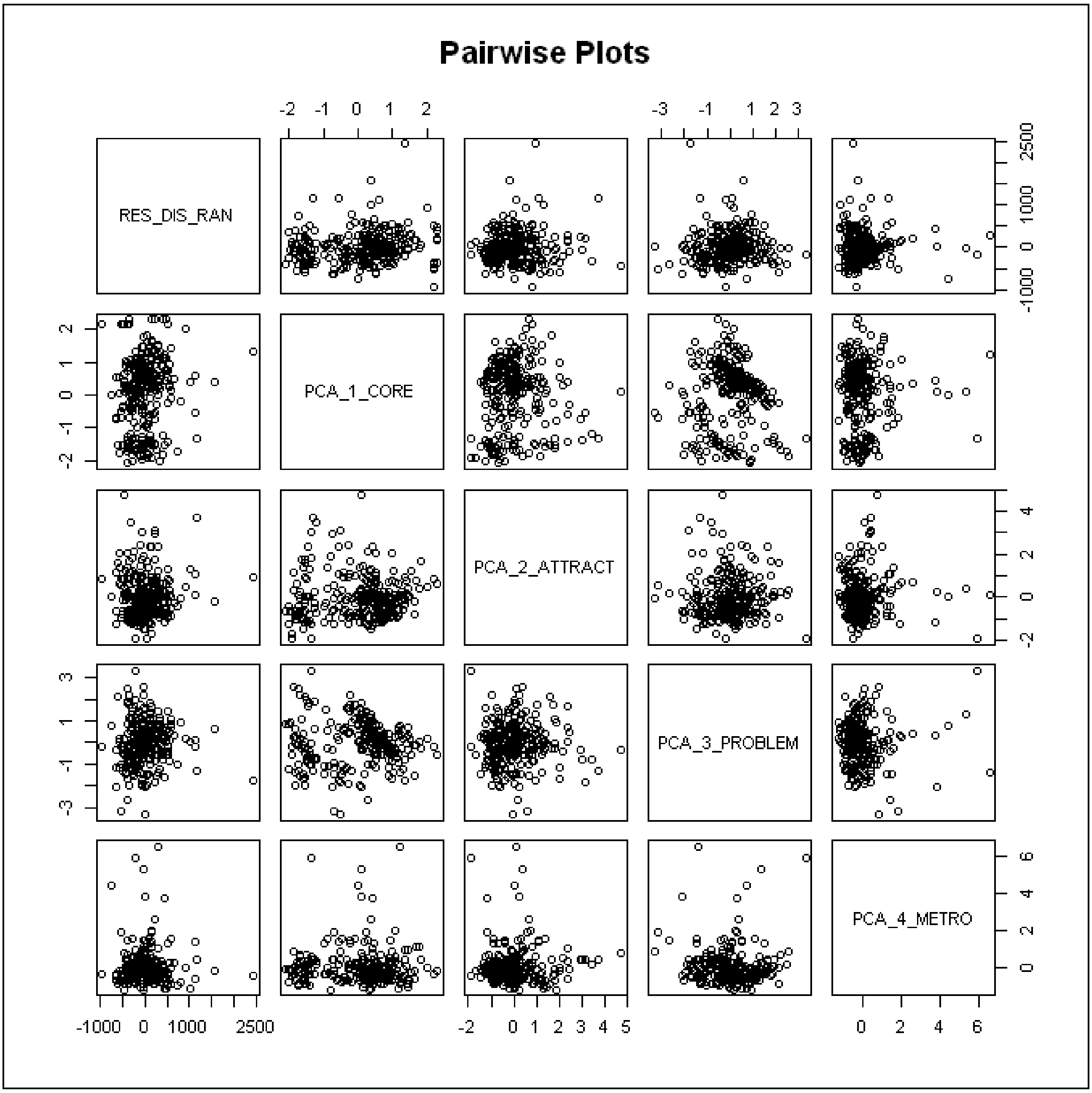


```
Data Type:        N

  1   2   3   4
 54  49 105  69
Categories found are:  1 2 3 4
-----------------------------------------------------------------
Variable:         PCA_1_CORE
Data Type:        R
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-2.0830 -0.5734  0.3008  0.1183  0.8610  2.3000
-----------------------------------------------------------------
Variable:         PCA_2_ATTRACT
Data Type:        R
    Min.   1st Qu.   Median     Mean  3rd Qu.     Max.
-1.92100 -0.68040 -0.04285  0.04166  0.61350  4.73600
-----------------------------------------------------------------
Variable:         PCA_3_PROBLEM
Data Type:        R
    Min.   1st Qu.   Median     Mean  3rd Qu.     Max.
-3.32800 -0.53000 -0.05204 -0.02480  0.52280  3.32100
-----------------------------------------------------------------
Variable:         PCA_4_METRO
Data Type:        R
    Min.   1st Qu.   Median     Mean  3rd Qu.     Max.
-1.24800 -0.54670 -0.24530 -0.01657  0.18830  6.52000
-----------------------------------------------------------------
Variable:         REG_TYP
Data Type:        N

 0 11 12 21 22 23 31 32
 1 31 21 26 39 16 54 89
Categories found are:  0 11 12 21 22 23 31 32
-----------------------------------------------------------------

Completed

+===============================================================+
### Bivariate data summaries                                  ###
+===============================================================+

Bivariate correlations and p-values
Correlations in the lower triangle and p-values in the upper
               RES_DIS_RAN  PCA_1_CORE PCA_2_ATTRACT PCA_3_PROBLEM PCA_4_METRO
RES_DIS_RAN    1.000000000  0.01766585    0.47513844   0.529419978   0.9277271
PCA_1_CORE     0.142468956  1.00000000    0.18735514   0.460380033   0.5985053
PCA_2_ATTRACT  0.043083622  0.07945161    1.00000000   0.752310689   0.8303181
PCA_3_PROBLEM  0.037944414 -0.04453479   -0.01904662   1.000000000   0.8948981
PCA_4_METRO   -0.005474645 -0.03177309   -0.01293360   0.007973589   1.0000000
-----------------------------------------------------------------
Hit <enter> to continue...
Hit <enter> to continue...
Hit <enter> to continue...
Hit <enter> to continue...
```

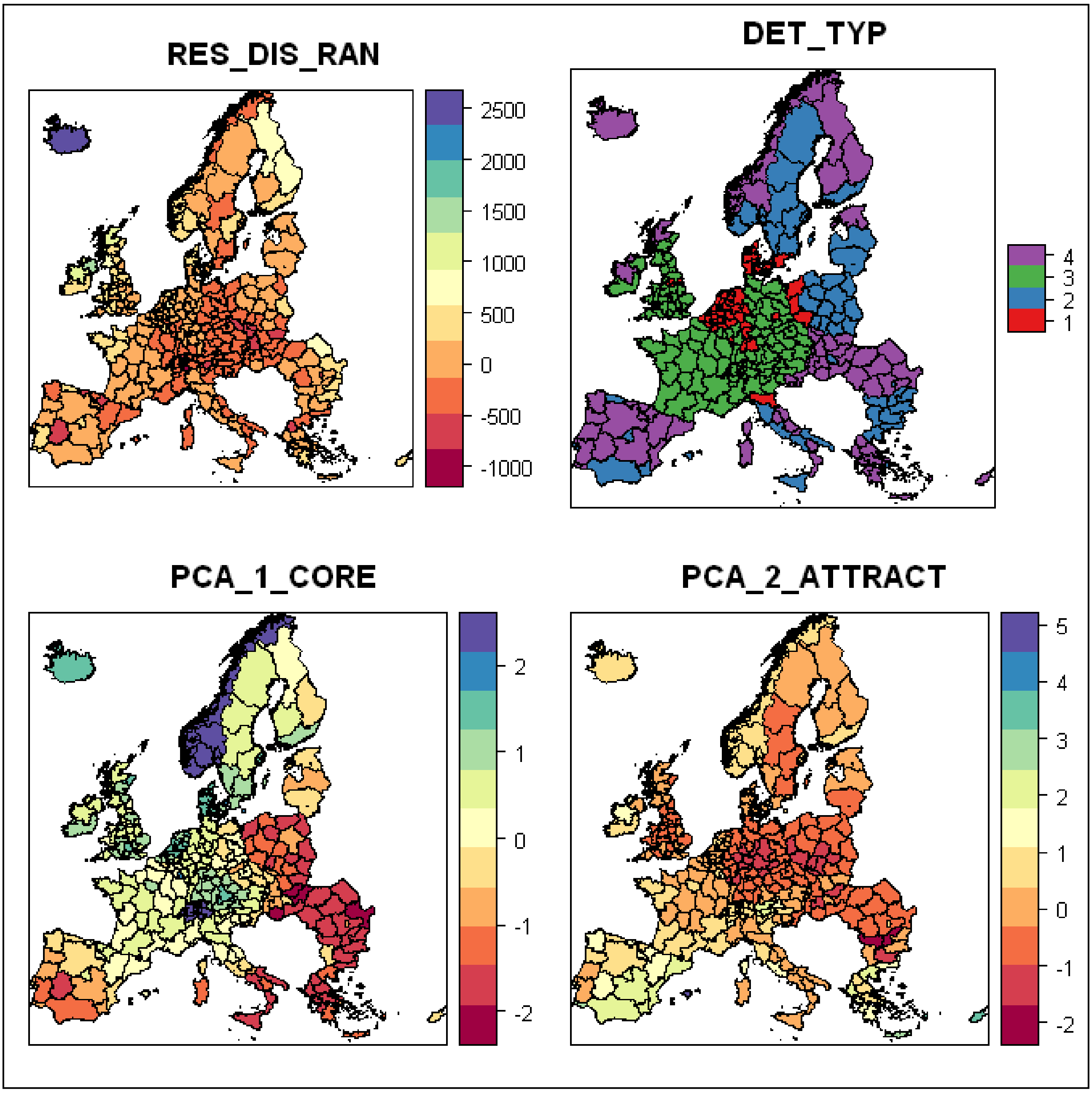
Barplots are used for the typology variables. In the case of REG_TYP the single value of zero seems unusual. This is not one of the allowable codes for this variable.

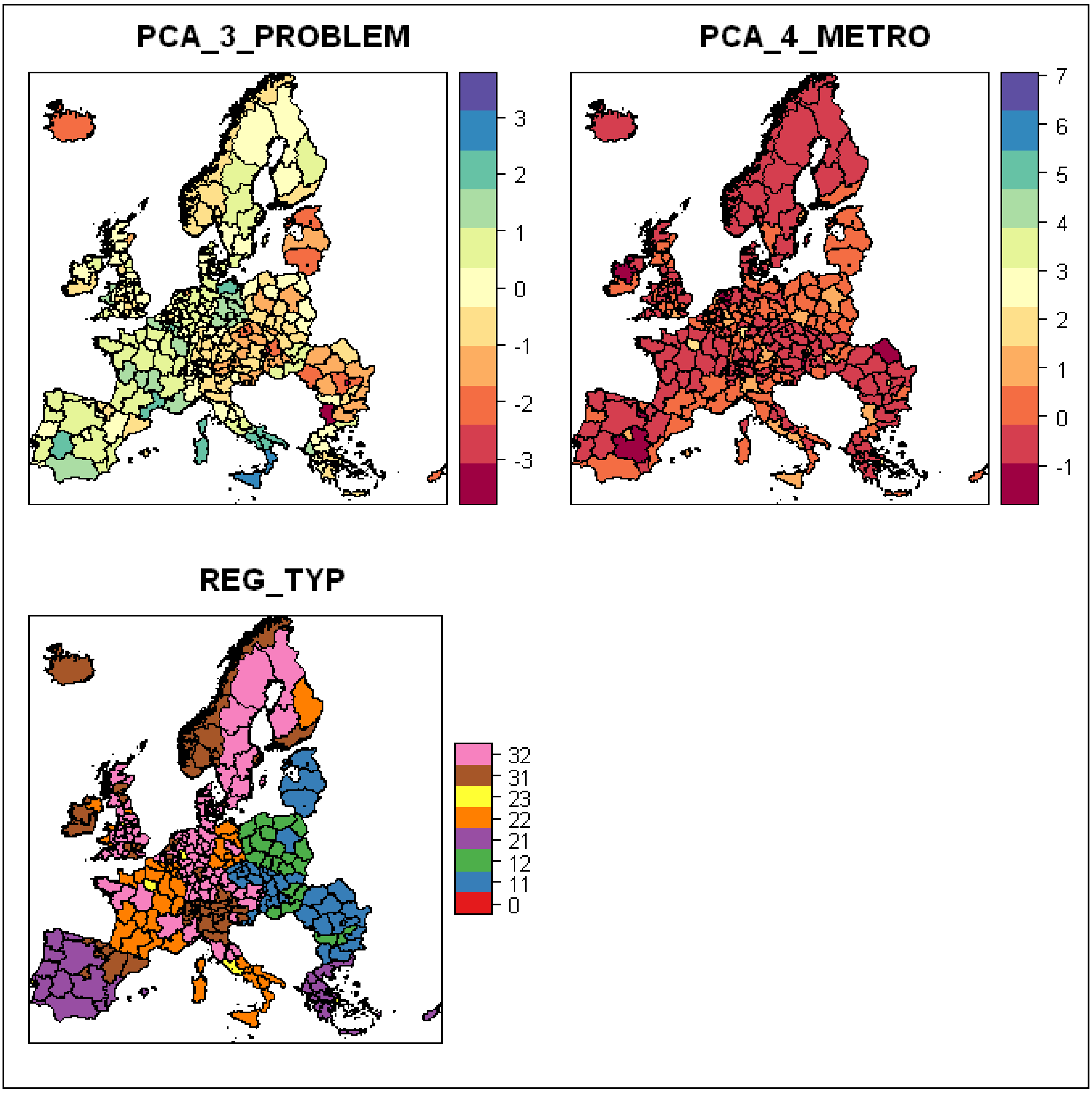


The scatterplot matrix presents plots for every combination of the variables in the multivariate list. These bivariate relationships will be examined again using the bagplot. The tendency for the fourth principal component to have uusually large values in its right tail will be noticed. The plots also help to highlight the orthogonality of the components. If any variable paris were strongly related this would be relatively clear from the plots – we might suggest to the data supplier that one of the variables is replaced by another choice.



The maps are produced automatically on groups of four. Those which display continous data have a standard key with 11 classes and a blue-red colour scale chosen from the ColorBrewer range of palettes. Those plots which display typology variables have another palette chosen from the ColorBrewer set – this highlights the visual differences between the classes. The shapefile has been reduced from its original extent by the removed of those regions for which there is no data.


RES_DIS_RAN
2500
2000
1500
1000
500
0
-500
-1000
DET_TYP
4
3
2
1
PCA_1_CORE
2
1
0
-1
-2
PCA_2_ATTRACT
5
4
3
2
1
0
-1
-2



Following the exploratory summaries, we begin the check for outliers. While for an individual test a region might be listed as a outlier, unless it is consistently identified across the univariate, spatial and multivariate tests as such, it is unlikely to be an outlier. We note first that there are a few regression residuals with large positive values. However, without details of the regression model itself, it would be unwise to drawn too many conclusions at this stage.

```
+===================================================================+
### Univariate Exception tests                                    ###
+===================================================================+
Ratio Data Check for Indicator:  RES_DIS_RAN

Regions with unusual boxplot data
NUTSCode     Value               NUTS Region Name
CH07         -957.766611449259   NA
GR42         1151.37411218726    ??t?? ???a??
IE01         1015.74777438423    Border, Midland and Western
IS00         2444.80020223689    NA
NL31         923.489129622647    Utrecht
PT17         1145.92495046647    Lisboa
```

```
UKM6        1138.36615331514    Highlands and Islands
UKN0        1569.51766679875    Northern Ireland
-------------------------------------------------------------------
Nominal data check for Indicator:  DET_TYP
Valid codes:  1 2 3 4

Frequency Tabulation
  1   2   3   4
 54  49 105  69

Anomalous data values found...
NUTSCode                   Value                      NUTS Region Name
ES70                       NA                         Canarias
PT30                       NA                         Região Autónoma da Madeira
------------------------------------------------------------------
```

The checks on the component scores follow. There are some problems with the representation of characters with diacritics. Note that the components which account for lower proportions of the variance show a tendency to have more distributional outliers.

```
Ratio Data Check for Indicator:  PCA_1_CORE
-------------------------------------------------------------------
Ratio Data Check for Indicator:  PCA_2_ATTRACT

Regions with unusual boxplot data
NUTSCode         Value            NUTS Region Name
CY00             3.13887          ??p??? / Kibris
ES53             4.73611          Illes Balears
GR22             3.44715          ????a ??s??
GR42             3.69832          ??t?? ???a??
GR43             3.02881          ???t?
PT15             2.93365          Algarve
-------------------------------------------------------------------
Ratio Data Check for Indicator:  PCA_3_PROBLEM

Regions with unusual boxplot data
NUTSCode               Value                  NUTS Region Name
BG41                   -3.32768               ??????????
DE30                   3.32121                Berlin
DE80                   2.17177                Mecklenburg-Vorpommern
ES43                   2.15609                Extremadura
ITF3                   2.20895                Campania
ITF6                   2.47295                Calabria
ITG1                   2.5697                 Sicilia
RO32                   -3.14105               Bucuresti - Ilfov
SK01                   -2.65745               Bratislavský kraj
-------------------------------------------------------------------
Ratio Data Check for Indicator:  PCA_4_METRO

Regions with unusual boxplot data
NUTSCode            Value               NUTS Region Name
AT13                4.40635             Wien
BE10                5.28288             Région de Bruxelles-Capitale
CZ01                3.80599             Praha
DE30                5.86853             Berlin
DE50                1.90913             Bremen
DE60                3.73206             Hamburg
DE71                1.43234             Darmstadt
FR10                2.00554             Île de France
GR30                1.54061             Att???
HU10                1.52142             Közép-Magyarország
```

```
ITE4                 1.42343            Lazio
PL12                 1.2921             Mazowieckie
PT17                 1.37259            Lisboa
RO32                 1.8552             Bucuresti - Ilfov
SK01                 1.41877            Bratislavský kraj
UKD3                 1.39142            Greater Manchester
UKG3                 1.52796            West Midlands
UKI1                 6.52024            Inner London
UKI2                 2.53976            Outer London
--------------------------------------------------------------------
```

The data check against the valid list of codes in the metadata reveals a code of 0 with a frequency of 1. This may well be an data entry error and can be quickly checked with the supplier.

```
Nominal data check for Indicator:  REG_TYP
Valid codes:  11 12 21 22 23 31 32

Frequency Tabulation
 0 11 12 21 22 23 31 32
 1 31 21 26 39 16 54 89

Anomalous data values found...
NUTSCode                   Value                      NUTS Region Name
ES70                       NA                         Canarias
MT00                       0                          Malta
PT30                       NA                         Região Autónoma da Madeira
--------------------------------------------------------------------
```

The spatial tests follow. The score runs from 1 – not an outlier on any test to 0 – an outlier on all tests. Those areas for which over half of the tests suggest outliers are listed.

```
+================================================================+
| Spatial exception tests                                        |
+================================================================+

RES_DIS_RAN   PCA_1_CORE    PCA_2_ATTRACT PCA_3_PROBLEM PCA_4_METRO

### Potential spatial exceptions  ###
NUTSCode       Value    NUTS Region Name
AT32           0.4      Salzburg
AT33           0.2      Tirol
BG32           0.4      ??????? ?????????
CH05           0.4      NA
CZ05           0.4      Severovýchod
CZ07           0.4      Strední Morava
CZ08           0.2      Moravskoslezsko
ES43           0.4      Extremadura
ES61           0.4      Andalucía
ITD1           0.4      Provincia Autonoma Bolzano/Bozen
PL22           0.4      Slaskie
PL51           0.4      Dolnoslaskie
PL52           0.4      Opolskie
RO31           0.4      Sud - Muntenia
RO41           0.4      Sud-Vest Oltenia
SK02           0.4      Západné Slovensko
```

The final set of tests are the mutlivariate tests. Again a score running from 1 (not an outlier) to 0 (almost certainly an outlier) summarises the results from the tests. It is noticeable that of the regions tested, most of the outliers are major cities (Vienna, Brussels, Berlin, London), or in the case of Iceland, the nation. The others include the Balearic Islands and the Aegean Islands.

```
+================================================================+
### Multivariate Exception tests  ###
+================================================================+
### Potential multivariate exceptions  ###
NUTSCode            Value               NUTS Region Name
AT13                3.2                 Wien
BE10                3.2                 Région de Bruxelles-Capitale
DE30                2.5                 Berlin
ES53                3.2                 Illes Balears
GR42                2.6                 ??t?? ???a??
IS00                2                   NA
UKI1                3.2                 Inner London
```

The resulting scores provide little evidence, apart from the apparently erroneous value of 0 for the REG_TYP variable, that there are any outliers of note in this dataset.

References

Ben-Gal I, 2005, Outlier detection, in Maimon O and Rockach L (eds) *Data Mining and Knowledge Discovery Handbook*, Dordrecht: Kluwer Academic Publishers

Besag JE, 1974, Spatial interaction and the statistical analysis of lattice systems, *Journal of the Royal Statistical Society*, Series B, 36, 192-236

Besag JE and Kooperberg C, 1995, On conditional and intrinsic autoregression, *Biometrika*, 82, 733-746

Bradley R and Haslett J, 1992, High-interaction diagnostics for geostatistyical models of spatially referenced data, *The Statistician*, 41, 371-380

Chambers R, Hentges A, Zhao X, 2004, Robust automatic methods for outlier and error detection., *Journal of the Royal Statistical Society*, Series A 167(2), 323-339.

Chandola V, Banerjee A and Kumar V, 2007, *Anomaly Detection: a survey*, Technical Report TR07-17, Minneapolis: Department of Computer Science, University of Minnesota

Chen D, Lu C-T, Kou Y and Chen F, 2008, On detecting spatial outliers, Geoinformatica, 12, 455-475

Cressie, N, 1993, *Statistics for Spatial Data*, revised edition, New York: Wiley

Cruz Ortiz M, Sarabia LA, Herrero A, 2006, Robust regression techniques: A useful alternative for the detection of outlier data in chemical analysis. *Talanta* 70, 499-512.

Dixon WJ, 1950, Analysis of extreme values, *Annals of Mathematical Statistics*, 21(4), 488-506

Dixon WJ, 1951, Ratios involving extreme values, *Annals of Mathematical Statistics*, 22(1), 68-78

Filzmoser P, Garret RG, and Reimann C, 2005, Multivariate outlier detection in exploration geochemistry, *Computers and Geosciences*, 31, 579-587

Filzmoser P, Maronna R and Werner M, 200u8, Outlier detection in high dimensions, *Computational Statistics and Data Analysis*, 52, 1694-1711

Filzmoser P, Ruiz-Gazen A and Thomas-Agnan C, 2012, Identification of local multivariate outliers, submitted for publication

Gower, JC, 1971, A general coefficient of similarity and some of its properties, *Biometrics*, 27, 857-874

Grubbs FE, 1950, Sample criteria for testing outlying observations, *Annals of Mathematical Statistics*, 21(1), 27-58

Hawkins RM, 1980, *Identification of Outliers*, London: Chapman and Hall

Ihaka R, Gentleman R, 1996, R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics* 5, 299-314.

Krige DG and Magri EJ, 1982, Studies of the effects of outliers and data transformations on variogram estimates for a base metal and a gold ore body, *Mathematical Geology*, 14(6), 557-564

Jackson DA, Chen Y, 2004, Robust principal component analysis and outlier detection with ecological data, *Environmetrics* 15, 129-139.

Kou Y, Lu C-T, Chen D, 2006, Spatial Weighted Outlier Detection, in Proceedings of the 2006 SIAM International Conference on Data Mining No. 614 2006.

de Leon and Carrière, 2002

Locantore N, Marron J, Simpson D, Tripoli N, Zhang J, Cohen K, 1999, Robust principal components for functional data, *Test*, 8, 1–73.

Liu H, Jezek KC, and O'kelly M, 2001, Detecting outliers in regularly distributed spatial data sets by locally adaptive and robust statistical analyssi and GIS, *International Journal of Geographical Information Science*, 15(8), 721-741

Pearson ES and Sekar CC, 1936, The efficiency of statistical tools and a criterion for the rejection of outlying observations, *Biometrika*, 28(3), 308-320

Penny KI and Joliffe IR, 2001, A comparison of multivariate outlier detection methods for clinical laboratory safety data, *The Statistician*, 50(3), 295-308

R Development Core Team, 2005, *R: a language and environment for statistical computing, reference index version 2.14.1*, R Foundation for Statistical Computing, Vienna, Austria. ISBN: 3-900051-07-0 URL: http://www.R-project.org

Rossi, RE, Mulla DJ, Journel AG, and Franz EH, 1992, Geostatistical tools for modelling and interpreting ecological spatial dependence, *Ecological Monographs*, 62(2), 277-314

Rousseeuw PJ, Ruts I, Tukey JW, 1999, *The* Bagplot: A Bivariate Boxplot*. The American Statistician* 53, 382–387.

Rousseeuw PJ, Debruyne M, Engelen S, Hubert M (2006) Robust and outlier detection in chemometrics. *Critical Reviews in Analytical Chemistry* 36, 221-242

Snedecor GW and Cochran WG, 1980, *Statistical Methods*, 7th edition, Ames: Iowa State University Press

Torgo, L, 2007, Resource bounded fraud detection, in Neves et al, (eds), *Progress in Artificial Intelligence*: 13th Portugese Conference on Artificial Intelligence, EPIA 2007, Lecture Notes in Artificial Intelligence, Heidelberg: Springer

Torgo, L, 2010, Data Mining with R, learning with case studies, Boca Raton: CRC Press

Tufte, ER, 1983, *The Visual Display of Quantitative Information*, Cheshire CT: graphics Press

van der Loo, MPJ, 2010, *Distribution based outlier detection for univariate data*, Discussion Paper 10003, Den Haag, Statistics Netherlands

Wall, MM, 2004, A close look at the spatial structure implied by CAR and SAR models, *Journal of Statistical Planning and Inference*, 121, 311-324

Wright DB and London K, 2009, *Modern Regression Techniques Using R*, London: Sage