Robust Principal Component and Factor Analysis in the Geostatistical Treatment of Environmental Data

Peter Filzmoser
Department for Statistics and Probability Theory
Vienna University of Technology, AUSTRIA

Abstract: In this paper we show the usage of robust multivariate statistical methods in geostatistics. A usual procedure to estimate the values of variables (e.g. geochemical variables) measured at certain points of a region is to apply geostatistical methods like Krige estimation (based on the estimation of variograms). Here we emphasize robust principal component and factor analysis for the preliminary investigation of the data to reduce the dimension. Geostatistical methods are applied afterwards to the estimated factor scores. The final results show the influence of certain combinations of variables in the considered region. Moreover, the estimated factor scores with the robust procedure indicate outlying observations in a much better way.

Keywords: Principal component analysis; Factor analysis; Robustness; Krige estimation; Variogram

1 Introduction

Environmental data are often measured at certain points of a region. We consider the case that at each point in the region, p variables have been measured. These variables may be chemical elements measured in soil. If we are interested in the distribution of the variables in the whole region, we could apply classical geostatistical methods which consider the spatial dependencies of the measured variables (see e.g. **?**). Spatial dependencies can be estimated with the *variogram* and the regional distribution is estimated using *Kriging*. The results can be shown in maps where the regional distribution of each variable is represented on its own map. Moreover, the standard error of the estimation can be calculated and may also be represented on a map.

A different question is to find the reasons for the regional distribution of the p variables. If we consider data describing the compound of the air at certain measurement points, it is interesting to find the reasons for the distribution of the given variables. Therefore we are not interested in maps which show the regional distribution for each variable separately, but we want to find combinations of the variables which reflect the cause of that distribution.

The above objective can be fulfilled by a first investigation of the variables with principal component analysis (PCA) (see e.g. **?**) or factor analysis (FA) (see e.g. **?**). By means of these multivariate methods the dimension will be reduced while keeping most of the information. PCA and FA perform a decomposition of the data matrix into a matrix of loadings which describes the connections between the variables and the new co-ordinate system, and a matrix of factor scores which consists of the variable values in the new co-ordinate system of the extracted factors.

Since both, PCA and FA are sensitive with respect to outliers, we emphasize robust methods. **?**) has shown a lot of possibilities to robustify classical multivariate methods, and he also applied these techniques to the analysis of geostatistical variables. Here we want to combine such methods with geostatistical tools. Section 2 treats a robust PCA method developed by **?**), and Section 3 is concerned with robust FA based on the MVE-estimator (**?**). The resulting factor scores are analyzed with the Krige estimation. For both methods examples with a practical data set are given.

2 Robust Principal Component Analysis

Classical PCA is sensitive to outlying observations. This can easily be seen since even the estimations of the covariance and correlation matrix are strongly influenced by outliers. As a consequence, the estimation of eigenvectors and eigenvalues is also strongly dependent on outliers in the data. Further errors might appear with the estimation of the mean vector which is necessary for both the centering of the data and the calculation of the classical covariance matrix. Therefore, we want to have a method which is not influenced by outliers, a *robust* method.

There are different "degrees" of robustness. A measure for the determination of the robustness of an estimator is given by the *breakdown value* (**?**). It is defined by the minimum proportion of contaminated data which causes the estimator to give arbitrary values.

In the following we describe a robust PCA method by **?**) which is based on the idea of *projection pursuit* (see e.g. **?**). Projection pursuit is a method for finding interesting structures in a p-dimensional data set following a certain criterion.

Classical PCA can be seen as a special case of projection pursuit. The criterion for finding a structure is the maximization of the variance of one-dimensional projections with the restriction of orthogonality to previously found directions.

For a better understanding of the robust method we first want to denote classical PCA with the help of the projection pursuit algorithm.

Let $\boldsymbol{X} = (\boldsymbol{x}_{1.}^\top, \ldots, \boldsymbol{x}_{n.}^\top)^\top$ be a data matrix with observation vectors $\boldsymbol{x}_{i.} \in I\!R^p$ $(i = 1, \ldots, n)$. We assume that the first $(k-1)$ eigenvectors $\widehat{\boldsymbol{\gamma}}_{.1}, \ldots, \widehat{\boldsymbol{\gamma}}_{.(k-1)}$ have been found. The k-th eigenvector will be defined as the maximum of the function

$$\boldsymbol{a} \quad \longrightarrow \quad S(\boldsymbol{X}\boldsymbol{a}) \tag{1}$$

under the restrictions

$$\boldsymbol{a} \perp \widehat{\boldsymbol{\gamma}}_{.1}, \ldots, \boldsymbol{a} \perp \widehat{\boldsymbol{\gamma}}_{.(k-1)} \qquad \text{and} \qquad \boldsymbol{a}^\top \boldsymbol{a} = 1\ .$$

S denotes the "classical" sample variance. A vector $\boldsymbol{a}$ which maximizes function (1) gives the new direction, namely the k-th estimated eigenvector. This means that we search step by step for a one-dimensional projection direction by maximizing the sample variance under certain restrictions.

With these considerations we now can describe the idea of robust PCA in a better way. Let us assume that the first $(k-1)$ projection directions $\widehat{\boldsymbol{\gamma}}_{.1}, \ldots, \widehat{\boldsymbol{\gamma}}_{.(k-1)}$ are already known. We define a projection matrix

$$\boldsymbol{P}_1 = \boldsymbol{I}_p \,, \qquad \boldsymbol{P}_k = \boldsymbol{I}_p - \sum_{j=1}^{k-1} \widehat{\boldsymbol{\gamma}}_{.j} \widehat{\boldsymbol{\gamma}}_{.j}^\top \,. \tag{2}$$

$\boldsymbol{P}_k$ corresponds to a projection onto the space spanned by the first $(k-1)$ projection directions. We are interested in finding a projection direction $\boldsymbol{a}$ which maximizes the function

$$\boldsymbol{a} \quad \longrightarrow \quad S(\boldsymbol{X}\boldsymbol{P}_k\boldsymbol{a}) = S\left(\begin{pmatrix} \boldsymbol{x}_{1.}^\top \\ \vdots \\ \boldsymbol{x}_{n.}^\top \end{pmatrix} \boldsymbol{P}_k \boldsymbol{a}\right) \tag{3}$$

under the restrictions

$$\boldsymbol{a}^\top \boldsymbol{a} = 1 \qquad \text{and} \qquad \boldsymbol{P}_k \boldsymbol{a} = \boldsymbol{a} \,. \tag{4}$$

The last restriction in Equation (4) ensures orthogonality to previously found projection directions.

The robustness of this method will be obtained by defining S as a robust measure of spread, e.g. as the *median absolute deviation* (MAD)

$$\text{medmed}(\boldsymbol{y}) = \ \underset{i}{\text{med}} |y_i - \ \underset{j}{\text{med}}\,(y_j)| \tag{5}$$

(breakdown value 50%).

There is an infinite number of directions we have to search for to maximize function (3). Therefore it is necessary to give an algorithm which approximates the solution.

For saving computation time we just search for projection directions belonging to the set

$$A_{n,k} = \left\{ \frac{\boldsymbol{P}_k(\boldsymbol{x}_{1.} - \widehat{\boldsymbol{\mu}}_n)}{\|\boldsymbol{P}_k(\boldsymbol{x}_{1.} - \widehat{\boldsymbol{\mu}}_n)\|}, \ldots, \frac{\boldsymbol{P}_k(\boldsymbol{x}_{n.} - \widehat{\boldsymbol{\mu}}_n)}{\|\boldsymbol{P}_k(\boldsymbol{x}_{n.} - \widehat{\boldsymbol{\mu}}_n)\|} \right\} \,. \tag{6}$$

$\widehat{\boldsymbol{\mu}}_n$ denotes the L_1-median defined by

$$\widehat{\boldsymbol{\mu}}_n = \ \underset{\boldsymbol{\mu} \,\in\, \mathbb{R}^p}{\text{argmin}} \sum_{i=1}^{n} \|\boldsymbol{x}_{i.} - \boldsymbol{\mu}\| \tag{7}$$

(**?**). The L_1-median has a breakdown value of 50%, is orthogonally equivariant, and is fast to compute.

This means that we search just for n directions determined by standardized vectors from the data points to the robust mean which are projected to a subspace. For large sample sizes we also could randomly select a subset of $A_{n,k}$.

2.1 Example

We consider a data set from geochemistry which was obtained during an investigation of a cooperation between the *Central Kola Expedition*, the *Geological Survey of Finland*, and the *Geological Survey of Norway (NGU)* (**?**)[1].

[1]See also Website *http://www.ngu.no/Kola*

This research community is concerned with the examination of the pollution in the so-called *Kola*-region, an area at the boundary of Norway, Finland, and Russia. At 617 irregularly spaced sites samples of soil have been taken. The samples include different layers depending on the depth. Our data set consists of measurement on *humus* which is the upper layer. The researchers measured the contents of 40 different chemical elements in every sample. So we have a data set with x- and y-coordinates of the 617 taken samples, and 40 variables.

6 of the 40 variables have been eliminated from our investigation because the concentration of these elements in the soil was very low. These values have been extremely inaccurate, and they probably would influence the results of our examination. The 34 remaining elements are Ag, Al, As, B, Ba, Bi, Ca, Cd, Co, Cr, Cu, Fe, Hg, K, La, Mg, Mn, Mo, Na, Ni, P, Pb, Rb, S, Sb, Sc, Si, Sr, Th, Tl, U, V, Y, and Zn. All these variables have been transformed to a logarithmic scale to give a better approximation to the normal distribution.

The 34 variables are analyzed with classical PCA. In order to put everything to a common scale we first standardized the variables to mean zero and variance one. The first principal component (PC) explains about 29% of the total variation, component 2 explains 14%, PC 3 has 11%. The first 6 PCs explain about 72% of the total variation.

We want to compare these values with robust PCA. Therefore, we first standardized the variables in a robust sense to median zero and median absolute deviation one (see Equation (5)). The results of the proportion of total variation for the robust PCs are: PC 1 10.2%, PC 2 7.9%, PC 3 6.4%. The first 6 robust PCs explain about 40% of the total variation which is much less than that obtained by classical PCA. The reason for this phenomenon is that there are outliers which strongly determine the direction of the classical PCs and therefore the proportion of total variation. Even if we compare this proportion between the first classical and robust PC (29% and 10.2%), we realize a big difference which is caused by outliers.

With PCA we decompose the data matrix into a matrix of loadings and a matrix of scores. The loadings of the first classical and robust PC are compared in Figure 1. We see that the classical and robust loadings of the first component are in most cases very different from each other. However, it is difficult to find interpretations for both the first classical and robust PC since a lot of variables contribute to the variation of the corresponding component.

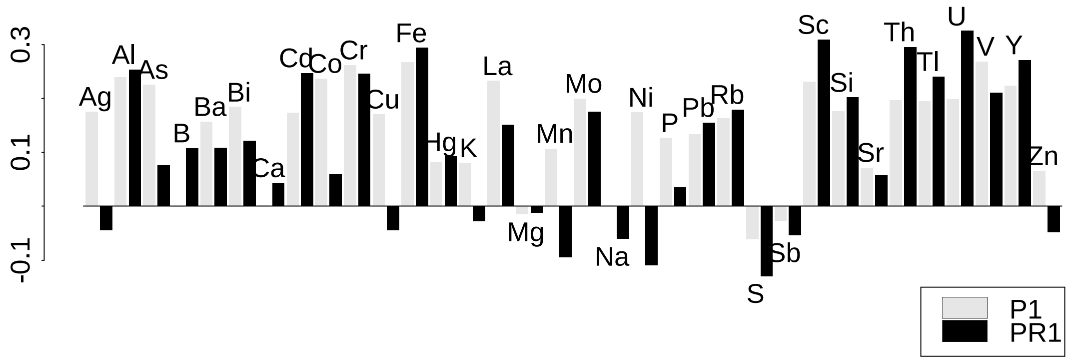

Figure 1: Loadings of the First Classical (P1) and Robust (PR1) Principal Component.

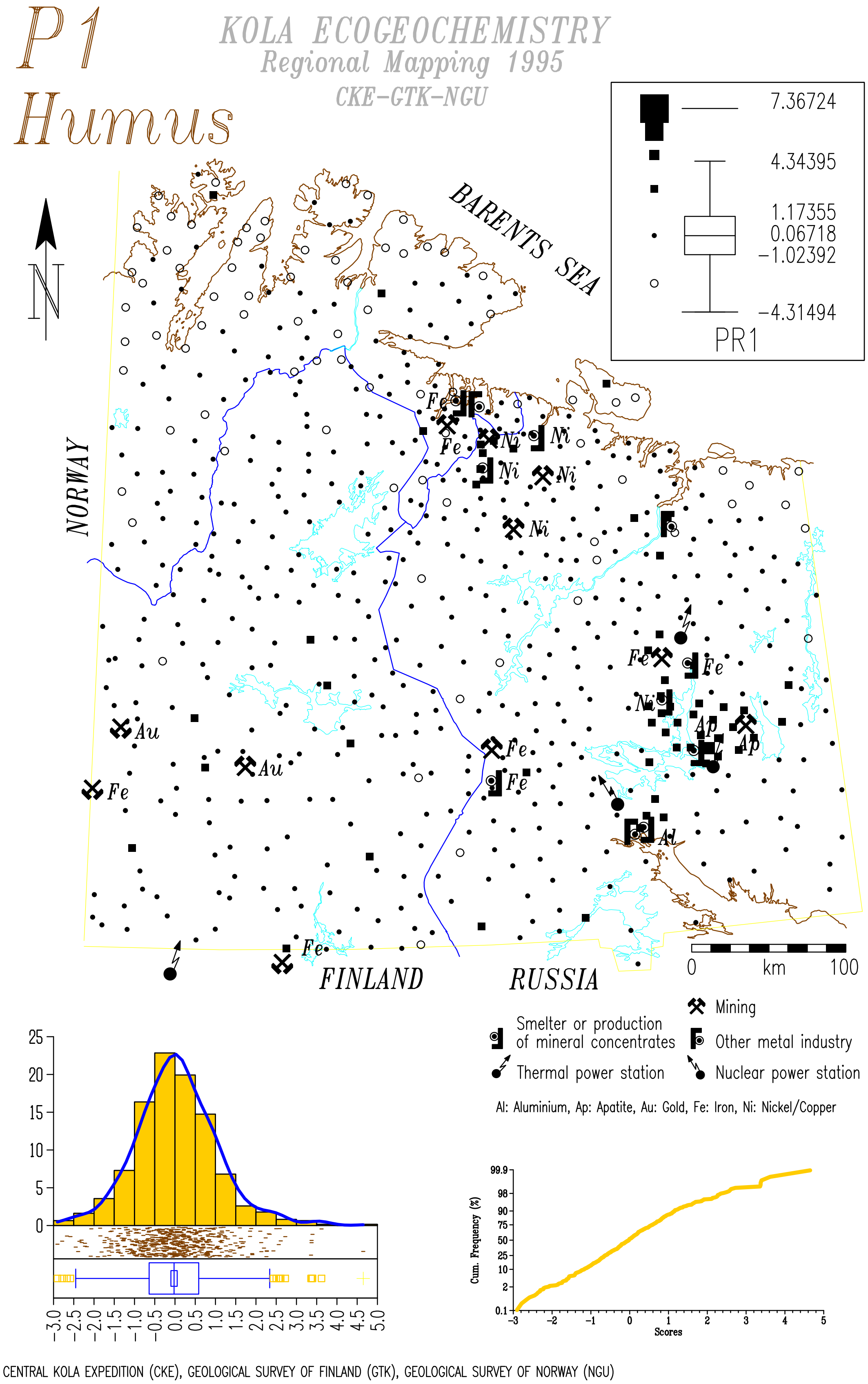

Figure 2: Scores of the First Classical Principal Component.

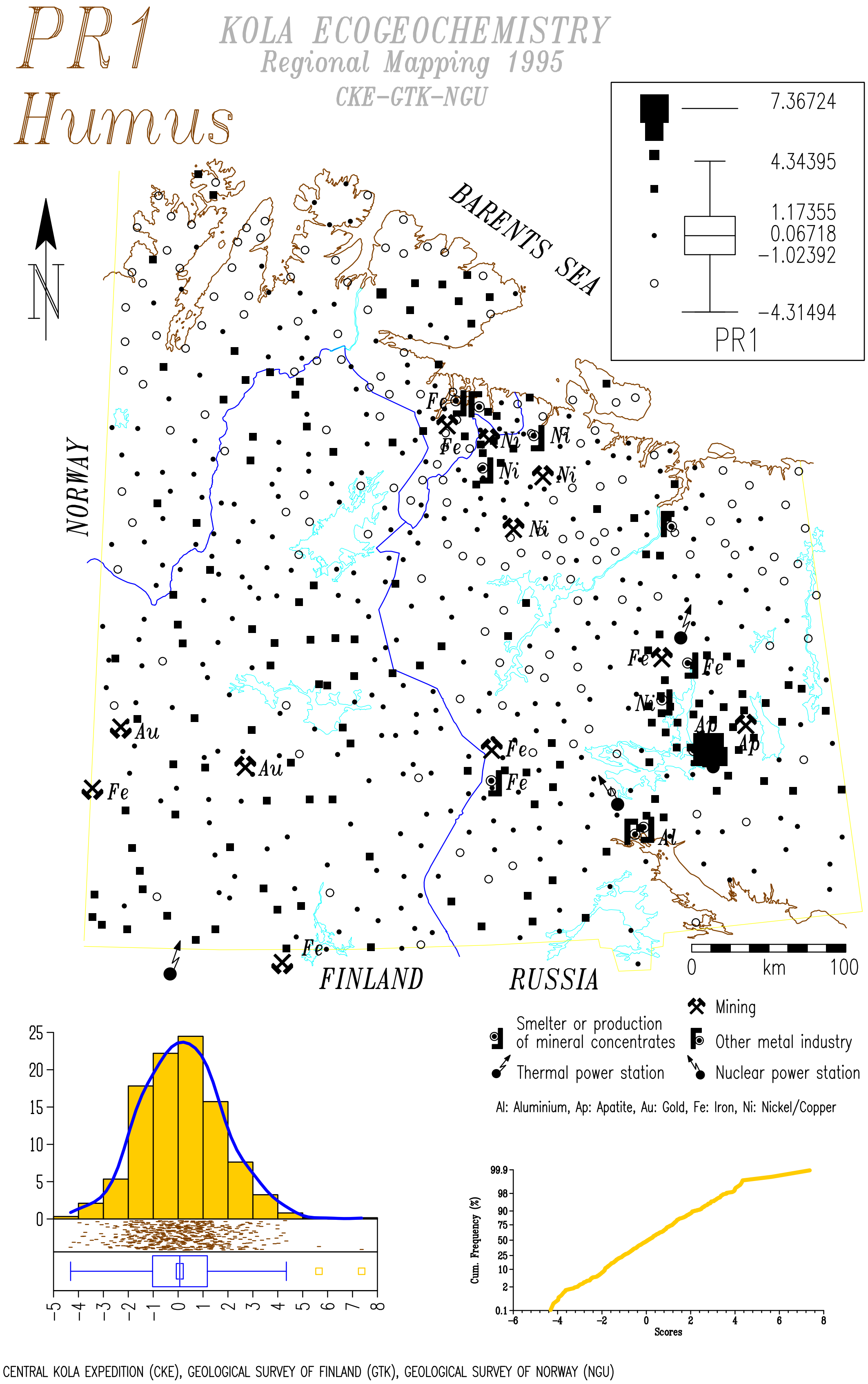

Figure 3: Scores of the First Robust Principal Component.

We also want to represent the scores of the first classical and robust PC graphically. In the bottom of Figures 2 and 3 there are some univariate graphics for the scores of the first classical and the first robust PC, respectively. The left picture shows a histogram, a density estimation, a univariate scattergram and a boxplot. The boxplot of the scores from classical PCA shows a lot of outliers which mainly determine the direction of this component. In comparison the boxplot of the robust scores shows just two outliers which, of course, do not determine the direction since for robust PCA we use a highly robust measure of scale.

The right lower diagrams show the cumulative frequencies of the scores in units according to the normal distribution. A straight line would indicate normally distributed scores.

The maps in Figure 2 and 3 are drawn with the computer program DAS (**?**). They show the region under consideration. Mining, smelting, metal industry, and power stations are marked with different symbols. At the 617 measurement points where samples have been taken we show the values of the scores. The scale of the scores has been taken from the boxplot of the robust scores (see legend in the upper right). We clearly see the location of the two outliers of the robust scores which are situated in an area with a lot of industry. If we compare the maps for classical and robust scores, the map for robust scores makes some areas with a lot of high (low) values visible like, e.g., in the south-west.

As mentioned before, it is difficult to find interpretations of the components since there is contribution of a variety of different variables. The next section is concerned with factor analysis which is also a technique to reduce the dimension with the aim of finding factors with interpretation.

3 Robust Factor Analysis

The factor analytical model is based on the assumption that what can be measured is but a realization of quantities in the background, which can not be measured directly. In many cases this assumption is realistic. Therefore it is of interest to determine whether quantities can be isolated from the variables observed which explain the correlation.

The main goal of factor analysis is the derivation of hypothetical quantities or factors from a set of variables observed. The factors are to be as simple as possible, and they should describe and explain the observations with sufficient precision (so that they can be interpreted).

Let $\boldsymbol{x} = (x_1, \ldots, x_p)^\top$ be a random vector with the variables $x_1, \ldots, x_p$. Let further $\boldsymbol{y} = (y_1, \ldots, y_p)^\top$ be the vector given by the transformation to mean zero and variance one (standardized variables),

$$\boldsymbol{y} = \boldsymbol{V_x}^{-1/2} \left[\boldsymbol{x} - \mathrm{E}(\boldsymbol{x})\right] ,$$

where

$$\boldsymbol{V_x} = \mathrm{Diag}(\mathrm{Cov}(\boldsymbol{x}))$$

is a diagonal matrix with the variances of $\boldsymbol{x}$.

We take the assumption that aside from an error term $\boldsymbol{e}$, the elements of $\boldsymbol{y}$ can be represented by a smaller number $k < p$ of unknown random variables $\boldsymbol{f} = (f_1, \ldots, f_k)^\top$

which are called *factors*. Then the classical k-factor model may be written in the way

$$\boldsymbol{y} = \boldsymbol{\Lambda} \boldsymbol{f} + \boldsymbol{e} \; . \tag{8}$$

$\boldsymbol{\Lambda} = [(\lambda_{ij})]$ is a $(p \times k)$-matrix with constant values. $\boldsymbol{\Lambda}$ is called *matrix of loadings* or *factor pattern*, and it describes the connections between factors and variables. The components e_i of the error term $\boldsymbol{e}$ are called *unique factors*.

In this model we assume

$$\begin{array}{ll} E(\boldsymbol{f}) = \boldsymbol{0} \; , \; \mathrm{Cov}(e_i, e_j) = 0 & \quad (i \neq j), \\ E(\boldsymbol{e}) = \boldsymbol{0} \; , \; \mathrm{Cov}(\boldsymbol{f}, \boldsymbol{e}) = \boldsymbol{O} \; . & \end{array}$$

The covariance matrix of $\boldsymbol{y}$ may therefore be expressed as

$$\begin{aligned} \mathrm{Cov}(\boldsymbol{y}) &= \boldsymbol{V_x}^{-1/2} \, \mathrm{Cov}(\boldsymbol{x}) \, \boldsymbol{V_x}^{-1/2} && (9) \\ &= \boldsymbol{\Lambda} \mathrm{Cov}(\boldsymbol{f}) \boldsymbol{\Lambda}^\top + \mathrm{Cov}(\boldsymbol{e}) \; . && (10) \end{aligned}$$

The essential step in factor analysis is the estimation of the matrix of loadings $\widehat{\boldsymbol{\Lambda}}$. Since in Equation (10) the number of unknown parameters would be too large, we assume for the present that the factors are orthogonal to each other. Then both the covariance matrix of the error term which is a diagonal matrix and the matrix of loadings can be estimated simultaneously (*maximum likelihood method*) or one after the other (*principal factor analysis)*. These estimations depend on the previously estimated covariance matrix of the standardized variables, which is directly derived by the covariance matrix of the variables $\boldsymbol{x}$ (Equation (9)). This means that all our parameter estimations depend on the initial estimation of the covariance matrix of the original variables.

We already mentioned in Section 1 that the classical estimation of a covariance matrix is extremely sensitive with respect to outliers. Even one "bad" observation can cause the estimation to become distorted. Therefore we recommend a robust estimation of the covariance matrix which, of course, could also be done in order to robustify PCA. Once we have a robust initial estimation of the covariance matrix of the original variables, all further parameter estimations are also robust.

There are different possibilities for a robust estimation of a covariance matrix. **?**) and **?**) developed the so-called M-estimator where the observations are weighted and afterwards the classical way of covariance estimation is done. The weight function should smoothly depend on the distance of the observations to the "center". A drawback of the M-estimator is the low breakdown value for high-dimensional data, namely at most $\frac{1}{p+1}$ where p is the number of variables.

A much better but for high dimensions computationally very expensive method is the MVE-estimator (*Minimum Volume Ellipsoid*) by **?**). The breakdown value of the MVE-estimator is the highest possible, namely 50%. The idea is to find an ellipsoid with minimum volume which covers half of the p-dimensional data points. The algorithm will be described in the following (**?**).

0. Initialize V_{min} (the unknown volume) with a large value.

1. Seek a subset of $p+1$ different observations (in *general position*) with the index set

$$J = \{i_1, \ldots, i_{p+1}\} \; .$$

2. Determine the arithmetic mean and the empirical covariance matrix for this subset:

$$\bar{\boldsymbol{x}}_J = \frac{1}{p+1}\sum_{i\epsilon J}\boldsymbol{x}_{i.}$$

$$C_J = \frac{1}{p}\sum_{i\epsilon J}(\boldsymbol{x}_{i.} - \bar{\boldsymbol{x}}_J)(\boldsymbol{x}_{i.} - \bar{\boldsymbol{x}}_J)^\top .$$

3. Calculate the radius of the ellipsoid (with $\bar{\boldsymbol{x}}_J$ and C_J) covering half of the points:

$$r_J^2 = \operatorname*{med}_i (\boldsymbol{x}_{i.} - \bar{\boldsymbol{x}}_J)^\top C_J^{-1}(\boldsymbol{x}_{i.} - \bar{\boldsymbol{x}}_J) .$$

4. Calculate the volume of the ellipsoid with

$$V_J = (\det(r_J^2 C_J))^{1/2} = (\det(C_J))^{1/2} r_J^p .$$

5. If V_J is less than V_{min} set $V_{min} = V_J$; go to 1.

6. The solution of the location estimation is $\bar{\boldsymbol{x}}_J$ and for the (pseudo-)covariance matrix

$$C = (\chi^2_{p,0.5})^{-1} r_J^2 C_J .$$

The final solution (with higher efficiency) is obtained by weighting the data and re-calculation of the usual covariance. The number of possible subsets in the algorithm is too high in most applications in order to get a reasonable computation time, and therefore a random selection can be done.

Note that the MVE-estimator yields both a robust location estimation and a robust estimation of the covariance matrix.

3.1 Example

We take the same data set as in the previous section and perform classical and robust FA. For robust FA we first have to calculate the MVE-estimator. Since the data matrix has 617 rows and 34 columns, this is a computationally very expensive task. However, a Pentium PC was able to compute the MVE-estimator with the program *Splus* within four hours. For both the classical and the robust method we chose the number of factors equal to six. In order to obtain a better interpretation for the factors, we rotated both the classical and the robust result of FA according to the oblimin-criterion (see **?**). The rotated loadings of the second factor are shown in Figure 4. If we just consider those variables with high contribution to the second factor, the interpretation of this factor might be *pollution*. This means, the variables with high absolute loadings on Factor 2 are elements typically reflecting pollution.

Like for PCA in the previous section, we can draw plots for the estimated classical and robust factor scores. Here we additionally want to estimate the spatial distribution of the second (robust) factor in the whole considered region which will be done by the Kriging estimator.

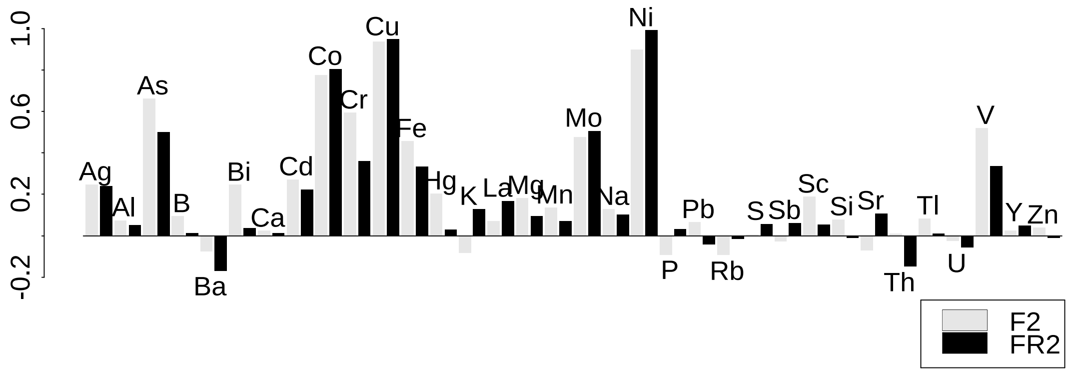

Figure 4: Loadings of the Second Classical (F2) and Robust (FR2) Factor.

On the basis of the estimated (robust) factor scores we calculated variograms for the east-west and north-south direction and estimate a variogram model. Variograms show the spatial dependencies of the scores. Afterwards, the whole region is divided into blocks, and in each block the Kriging estimator is applied. We obtained an estimation of the scores in each block and additionally an estimation of the Kriging error. These calculations have been done by the computer program GEOSAN (**?**). The results for the second classical and robust factor scores are shown in Figure 5 and 6, respectively. For both maps the Kriging results are scaled according to the given quantiles of the robust factor scores (see legend in the top right).

The comparison of Figures 5 and 6 shows in general an over-estimation in the west and an under-estimation in the east with the classical method. There are two areas with high pollution visible: one at the border of Norway and Russia and the other in the center of the considered part of Russia.

4 Conclusions

Principal component analysis and factor analysis are multivariate statistical methods which reduce the dimension. The components or factors obtained by these techniques are certain combinations of the original variables. One aim of FA is to obtain such combinations which give an interpretation. Especially for environmental data we are often interested in finding out quantities which describe a certain behavior, and therefore FA should be the method of choice.

If there are outliers in the data set, robust versions of PCA and FA should be considered. Although robust estimators require in general more computational effort, we have shown that even for high-dimensional data sets this procedure is still feasible. Results are not influenced by outlying observations, and they are a good basis for further investigations like geostatistical methods.

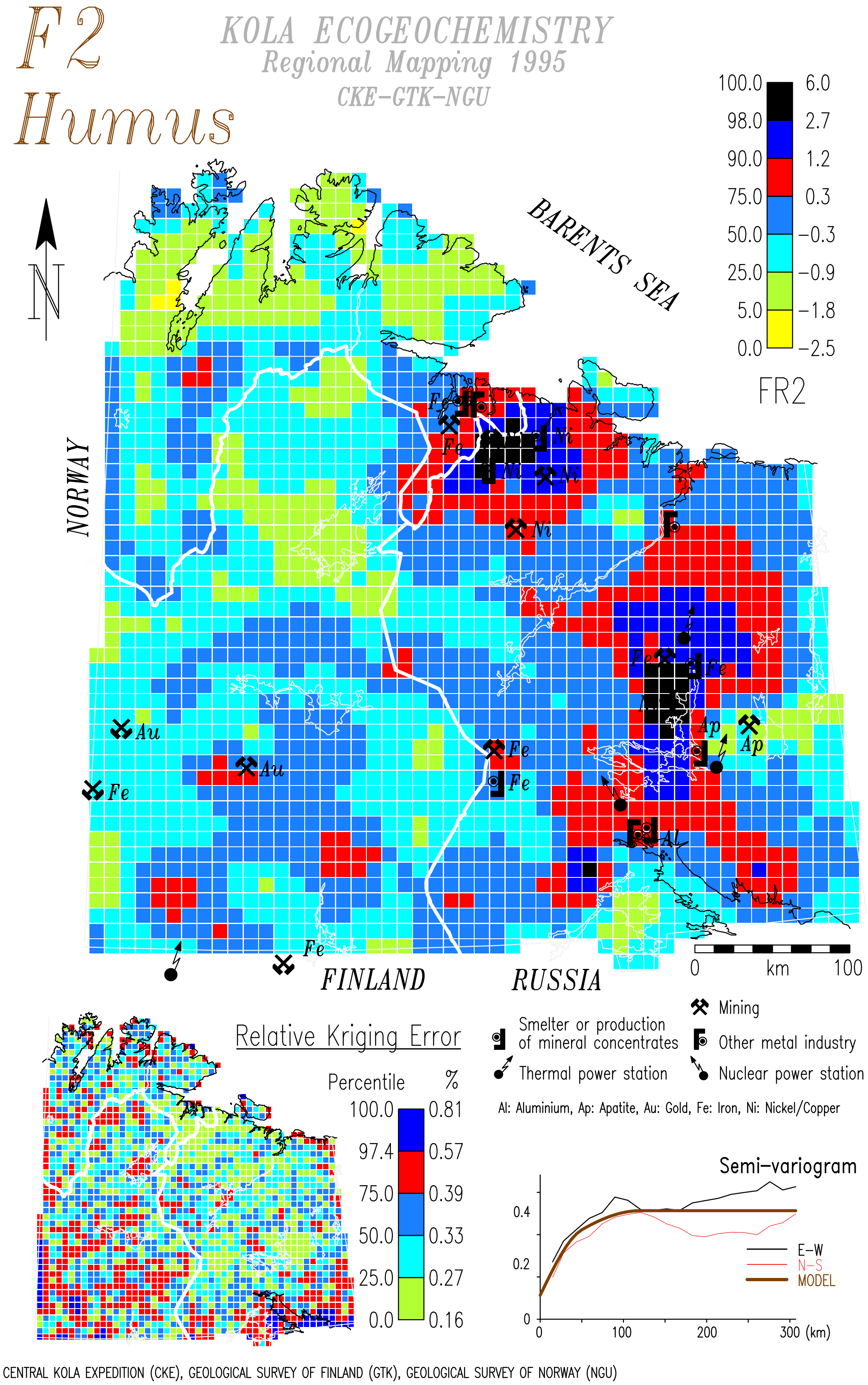

Figure 5: Krige-Results of the Second Classical Factor.

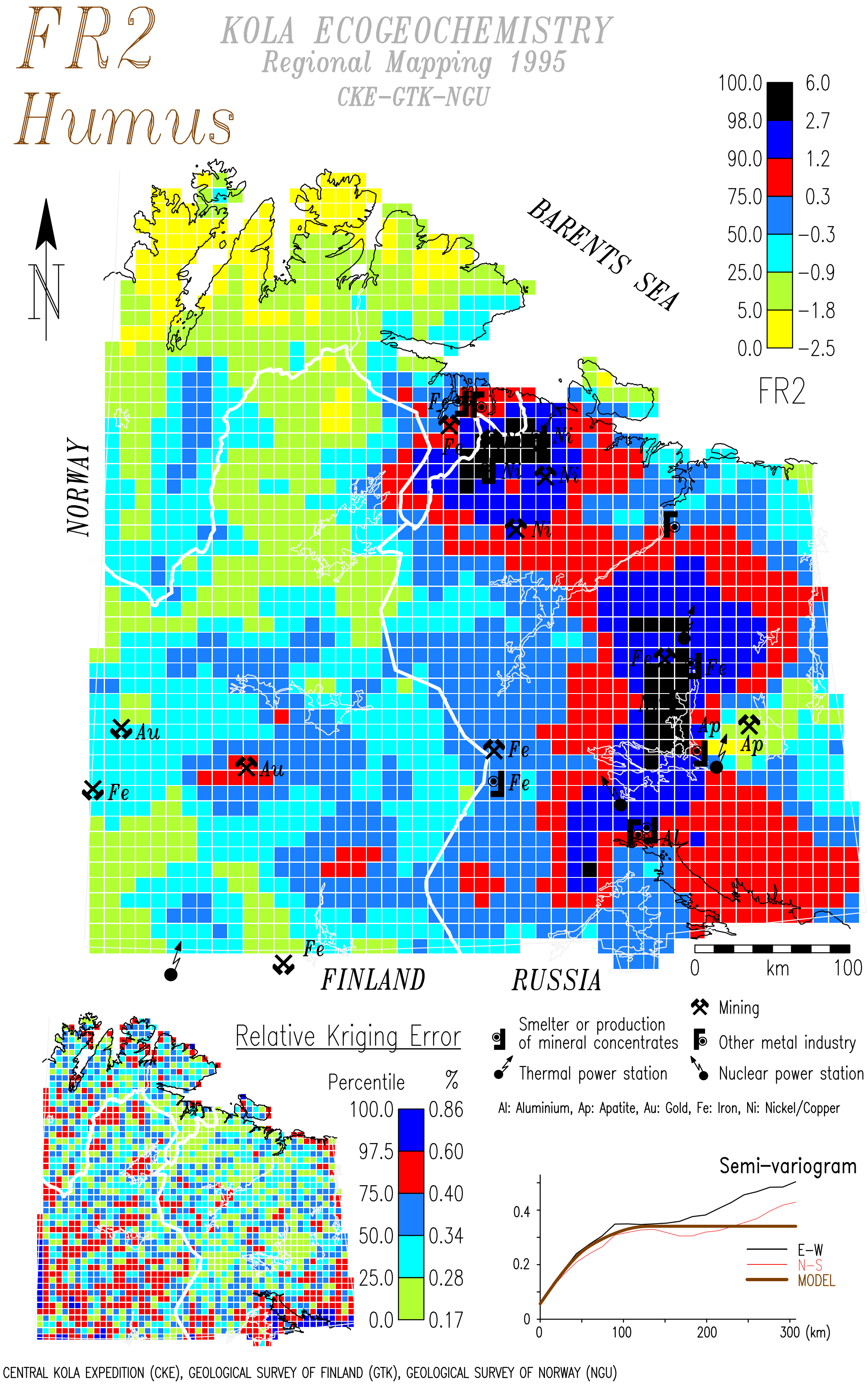

Figure 6: Krige-Results of the Second Robust Factor.

The MVE-estimator used here as a basis for FA (Section 3) could have also been applied to PCA. However, if the number of observations is in the magnitude of the number of variables or even smaller, the MVE-estimator cannot be further used (singularity). For this case, the method described in Section 2 is advisable. Besides the robust aspect there is another big advantage of robust to classical PCA: Since the components are estimated one after the other, we can stop the procedure after the calculation of a desired number of components instead of calculating all components.

Acknowledgment

The author wishes to thank Prof. Rudolf Dutter for placing the software products DAS and GEOSAN including several macros at his disposal, for the numerous hours of joint work with the computer, and for the help with geostatistical questions.

The Geological Surveys of Finland and Norway and Central Kola Expedition are thanked for making the data available for this investigation.