Modelling of Water Pollution in Urban Areas with GIS and Multivariate Statistical Methods

L. Matejicek

Charles University in Prague, Faculty of Science
Institute for Environmental Studies
Benatska 2, Prague, 12801
Czech Republic
lmatejic@natur.cuni.cz

Abstract: Integrating modelling of water pollution and GIS enables the connection of environmental process models with geospatial data describing the physical environment. Time series of water pollution data at monitoring profiles are used to complement the spatial database and to interpret the results of data analysis. In this case, the multivariate statistical methods provide an exploratory environment for data analysis and an indication of seasonal changes in the framework of surface water pollution. In addition to a wide range of useful multivariate methods, principal component analysis (PCA) and factor analysis (FA) are used to differentiate seasonal water pollution at monitoring profiles. Integrating modelling of water pollution is demonstrated for a river basin in the urban area of Prague. Data series from long-term measurements (25 years; seasonal measurements for PCA and FA in the period 2001-2004) are used to study the variability of water quality parameters. Consecutively, the PCA and FA are carried out to identify seasonal deviations originating from the time series of water temperature, pH, conductivity, suspended solids, nitrates, phosphates, BOD, COD, etc. The graphs focused on PCAs' loadings and FAs' biplots show the standard data outputs. The data from the multivariate seasonal exploratory analysis are transported into GIS to map the changes of FAs' loadings. The maps of changes are then used to estimate the observed seasonal strengths of the given processes that include simultaneous changes in water pollution parameters. As an example, the seasonal strength of NO_3 is mapped and compared together with other strengths of remaining parameters. The significant changes of FAs' loadings are observed between the winter seasons and the summer seasons (2001-2004), which is in correspondence with the original data. The described exploratory tools are developed to support decision-making processes in the framework of water pollution management.

Keywords: multivariate analysis; GIS; urban environment; surface water pollution

1. INTRODUCTION

The exploration of water pollution in urban areas represents a special case of research focused on water pollution. Due to the very high level of human interference with natural processes, all environmental phenomena must be considered in much smaller temporal and spatial scales than in rural areas. The essential differences with respect to methodology and data management mean that data collected by national meteorological services are seldom adequate for urban studies. Extended data collection has to be provided to deliver data on small spatial scales and short time resolutions. Locally collected data require long time periods before the amount of data is sufficient for meaningful exploratory applications. Physical properties and chemical composition of surface water systems in urban areas include many different types of soils that are heterogeneous and heavily disturbed. In addition to this heterogeneity, growing cities are constantly complemented with new elements. That is why the physical and chemical changes become more complex. A wide range of water pollution models and environmental analyses has been developed in the past decades. Some water pollution models have become a part of Geographical Information Systems (GIS). This integration brings the possibility to use water pollution analysis more efficiently [Maidment, 2002]. The data analysis of water pollution together with GIS is used in some developed countries on a municipal level for support of decision making systems, which can be

applied in real time. Thus, modelling of water pollution in urban areas is becoming an increasingly important tool for the assessment of environmental impacts. Urban water pollution will have an increasing influence on the sustainability of human societies. Growing urban populations and urban areas bring significant changes in physical properties of the land surface. Reinforced urban surfaces and channelling of natural streams result in fast runoff with high peak flows, which causes disastrous water pollution effects on the whole river basin downstream of the city. Thus, meaningful water pollution management cannot exist without monitoring networks extended by other surface data, calculation methods and exploratory techniques [Matejicek, 2002].

2. URBAN WATER POLLUTION DATA

Modelling of water pollution with GIS takes into account the huge amount of measurements and spatial characteristics. The measurements mostly represent time series of climatologic and hydrologic conditions, physical and chemical parameters of the river basins, estimates of water pollution from point and non-point sources, and potential ecological accidents. In the case of GIS, the spatial characteristics are derived from digital map layers (topographic maps, thematic maps focused on hydrological, soil and geochemical phenomena). In addition to the standard two-dimensional map layers, other spatial data can be incorporated into digital terrain models optionally extended by vegetation and buildings, and aerial images or satellite scenes.

In the case of this paper, water pollution measurements are based on samples taken regularly from the surface water network of the urban area of Prague. The investigated parameters (water temperature, water flow, conductivity, suspended solids, pH, oxygen saturation index, biological oxygen demand, chemical oxygen demand, nitrate, ammonia, phosphate, chloride, sulphate, calcium, magnesium, and coliform bacteria) have been measured for nearly 25 years. Due to successive changes in methodology, the various time interval settings during the period, and incompleteness of some data especially from the first years, only four years (2001-2004) of observations were selected for the present study. The location of the sampled water profiles are illustrated in schematic view in Figure 1. The data originated from 16 observation points, which are located on 4 major subsidiary water streams. The selected water streams with their observation points represent just a part of the whole monitoring network. These selections were made to identify the dominant water pollution.

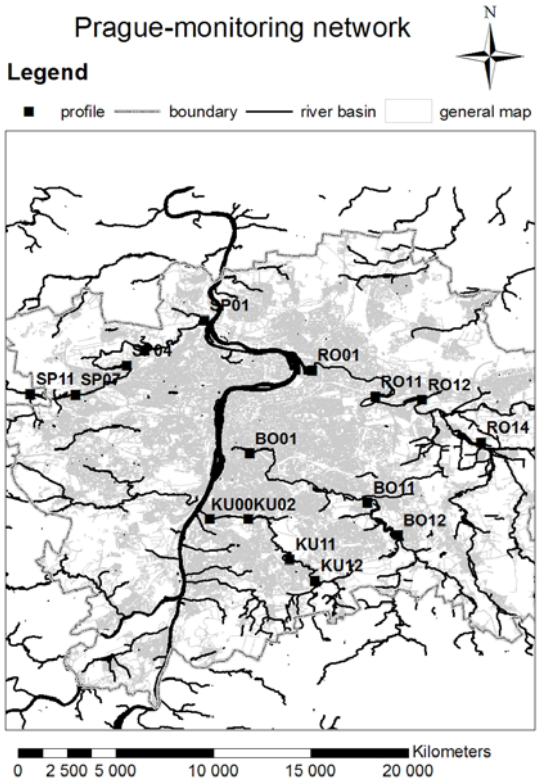


Figure 1. Monitoring network in the river basins.

3. METHODS FOR DATA PROCESSING

In the case of this paper, the multivariate statistical methods are focused on classification of the data collection taken from samples located on the surface water network in Prague. Principal component Analysis (PCA) and factor analysis (FA) are used to reduce the number of variables and to identify structure in the relationships among variables. The input data are standardised using the average and the standard deviation to remove different physical dimensions. By scaling all measured values with the standardisation, the non-dimensional total variance is defined as the sum of variances arising from the original variables. After processing data by PCA, a smaller number of the uncorrelated variables (principal components), which cover the majority of the variability, are found. In the framework of geometrical interpretation, the principal components are represented by orthogonal vectors, where the first major principal components course in the direction of the dominant data variability. In the case of FA, the orthogonal system can be rotated to find a more realistic interpretation of the principal axis.

These introduced techniques are just a part of multivariate statistical methods. The application of

PCA as a tool for water quality evaluation and management is described by Parinet [2004]. A case study focused on process identification by PCA in the framework of river water-quality data is represented in a paper written by Petersen [2001] and consecutively by Callies [2005]. For the river basins in the urban regions, some part of the variability can be explained by considering the external meteorological forcing, which is mainly related to the variations of water temperature and runoff. The organic compounds are highly related to biological activities, which are also dependent on seasonal cycles. Thus, two leading patterns of relations can be obtained, one discharge-dependent and the other caused by biological activities. But the processes in the urban environment are much more complex than in rural areas, which can be explored, in addition to multivariate statistics, by dynamic models [Matejicek, 2003]. Due to the heterogeneity of urban areas, the standard erosion models can be applied only partially, which means that the classification of river basin areas is used for just an approximate relationship together with the measured data. Thus, the multivariate statistical techniques are used to explore basic relationships that can be finally studied by dynamic models.

4. USING MULTIVARIATE STATISTICS

An overview of the input data for multivariate exploratory methods is illustrated in Figure 2. After the data standardisation, PCA and FA ("varimax rotation") are carried out separately for six seasons (January, March, May, July, September and November). The selection is subordinated to the available data and to the expected differences in biological activities, which appears to be the dominant source of qualitative changes. The results in Figure 3 represent PCAs' loadings and FAs' bi-plot diagrams. In all cases of the FAs' bi-plots, more than 50% of the variability is covered by two factors. To compare differences in factor loadings in each season (January, March, May, July, September, November), the factor loadings are shown together with coordinate control points ([0; 1], [1; 0], [-1; 0], [0; -1]) in the attached map schemas, Figure 4. To illustrate the changes of the factor loadings in a more transparent way, the points ranking to each specific parameter (water temperature, water flow, conductivity, suspended solids, pH, oxygen saturation index, biological oxygen demand, chemical oxygen demand, nitrate, ammonia, phosphate, chloride, sulphate, calcium, magnesium, and coliform bacteria) are interconnected by lines in the framework of the GIS project. As an example, the seasonal strength of NO_3 is mapped and compared together with other strengths of remaining parameters. The significant changes of FAs' loadings are observed between the winter seasons and the summer seasons (2001-2004), which is also in correspondence with the original data.

5. CONCLUSIONS

In spite of the fact that all data are managed by GIS (ArcGIS 9.x), multivariate statistical methods (PCA and FA) are carried out by a standalone statistical program (SPlus 6.1). But the idea is to present a compact software application for the decision-making processes in the framework of water pollution management. Thus, the proposed tools are intended to be integrated into various case tools as a standalone application. In the case of the presented paper, the long-term observations used are used for testing of developed tools, which nowadays assist in the interpretation of observed processes.

6. ACKNOWLEDGEMENTS

The authors wish to thank the authorities of the capital Prague and the Institute of Municipal Informatics of Prague for providing the data of the river basin and map layers of the urban areas.

7. REFERENCES

Callies, U., Interaction structures analyzed from water-quality data, *Ecological Modelling, 187*, 475-490, 2005.

Maidment, D.R., *ArcHydro: GIS for water resources*, ESRI, 220 pp., Redlands, 2002.

Matejicek, L., Environmental Modelling in Urban Areas with GIS, paper presented at the 1st Biennial Meeting of iEMSs, Lugano, Switzerland, June 24-27, 60-65, 2002.

Matejicek, L., L. Benesova and J. Tonika, Ecological modelling of nitrate pollution in small river basins by spreadsheets and GIS, *Ecological Modelling*, 170, 245-263, 2003.

Parinet, B., A. Lhote and B. Legube, Principal component analysis: an appropriate tool for water quality evaluation and management-application to a tropical lake system, *Ecological Modelling, 178*, 295-311, 2004.

Petersen, W., L. Bertino, U. Callies and E. Zorita, Process identification by principal component analysis of river-quality data, *Ecological Modelling, 138*, 193-213, 2001.

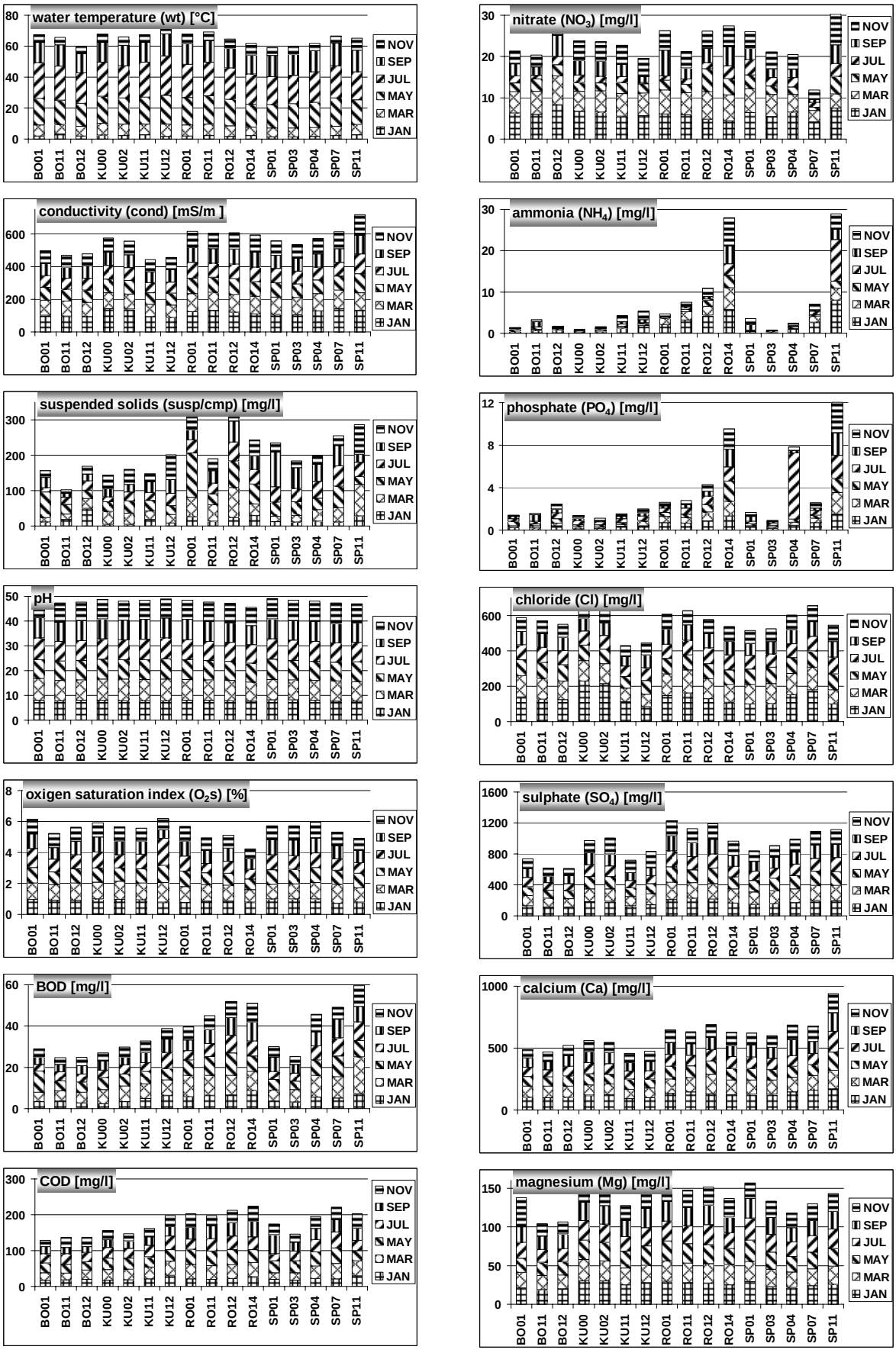
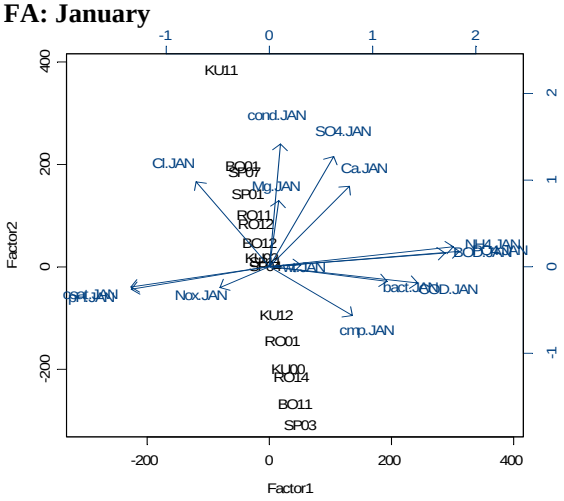


Figure 2. The pre-processed measured data.

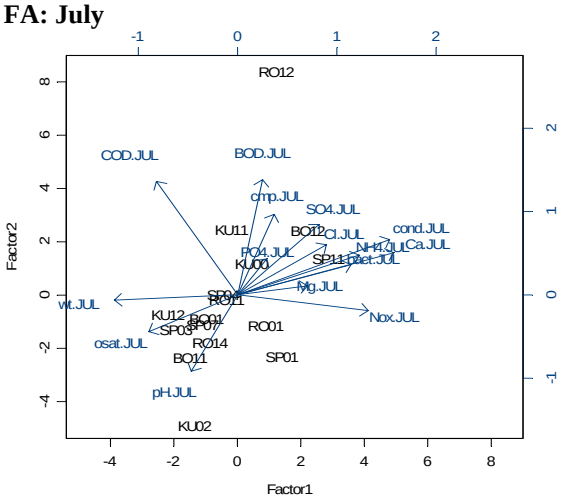
JANUARY	C1	C2	C3	C4	C5	C6
Proportion of variance	0.39	0.22	0.11	0.08	0.06	0.05
Cummulative proportion	0.39	**0.62**	0.73	0.81	0.87	0.92
MARCH	**C1**	**C2**	**C3**	**C4**	**C5**	**C6**
Proportion of variance	0.46	0.17	0.14	0.09	0.05	0.03
Cummulative proportion	0.46	**0.63**	0.77	0.86	0.91	0.95
MAY	**C1**	**C2**	**C3**	**C4**	**C5**	**C6**
Proportion of variance	0.34	0.22	0.13	0.11	0.07	0.05
Cummulative proportion	0.34	**0.56**	0.69	0.80	0.86	0.91

JULY	C1	C2	C3	C4	C5	C6
Proportion of variance	0.42	0.17	0.11	0.09	0.08	0.04
Cummulative proportion	0.42	**0.60**	0.71	0.80	0.88	0.92
SEPTEMBER	**C1**	**C2**	**C3**	**C4**	**C5**	**C6**
Proportion of variance	0.35	0.20	0.15	0.11	0.06	0.04
Cummulative proportion	0.35	**0.55**	0.71	0.82	0.88	0.92
NOVEMBER	**C1**	**C2**	**C3**	**C4**	**C5**	**C6**
Proportion of variance	0.35	0.20	0.15	0.11	0.06	0.04
Cummulative proportion	0.35	**0.55**	0.71	0.82	0.88	0.92

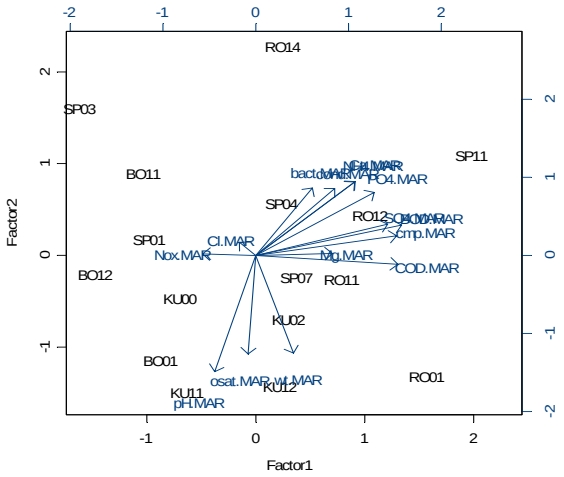
FA: January



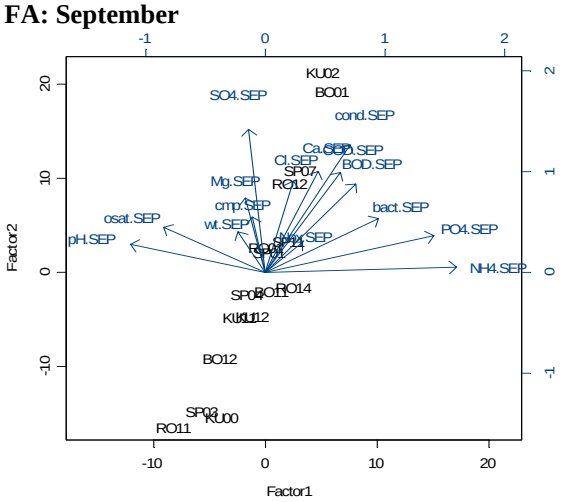
FA: July



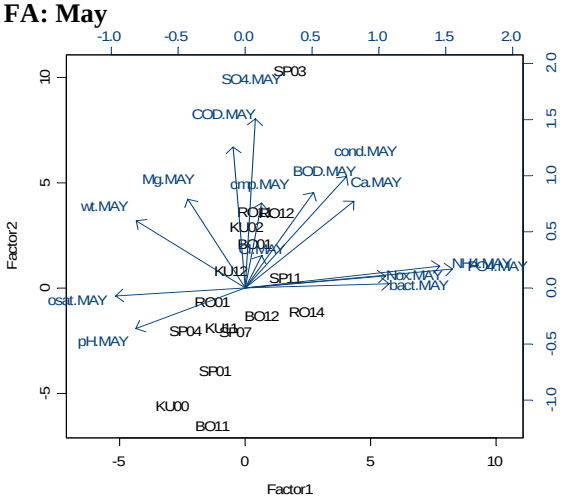
FA: March



FA: September



FA: May



FA: November

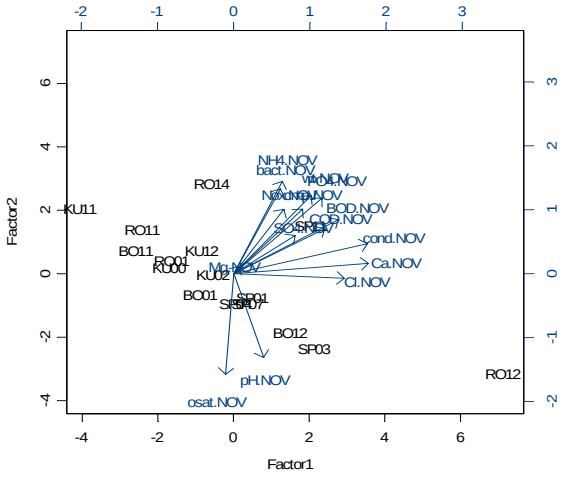


Figure 3. PCA and FA seasonal outputs.

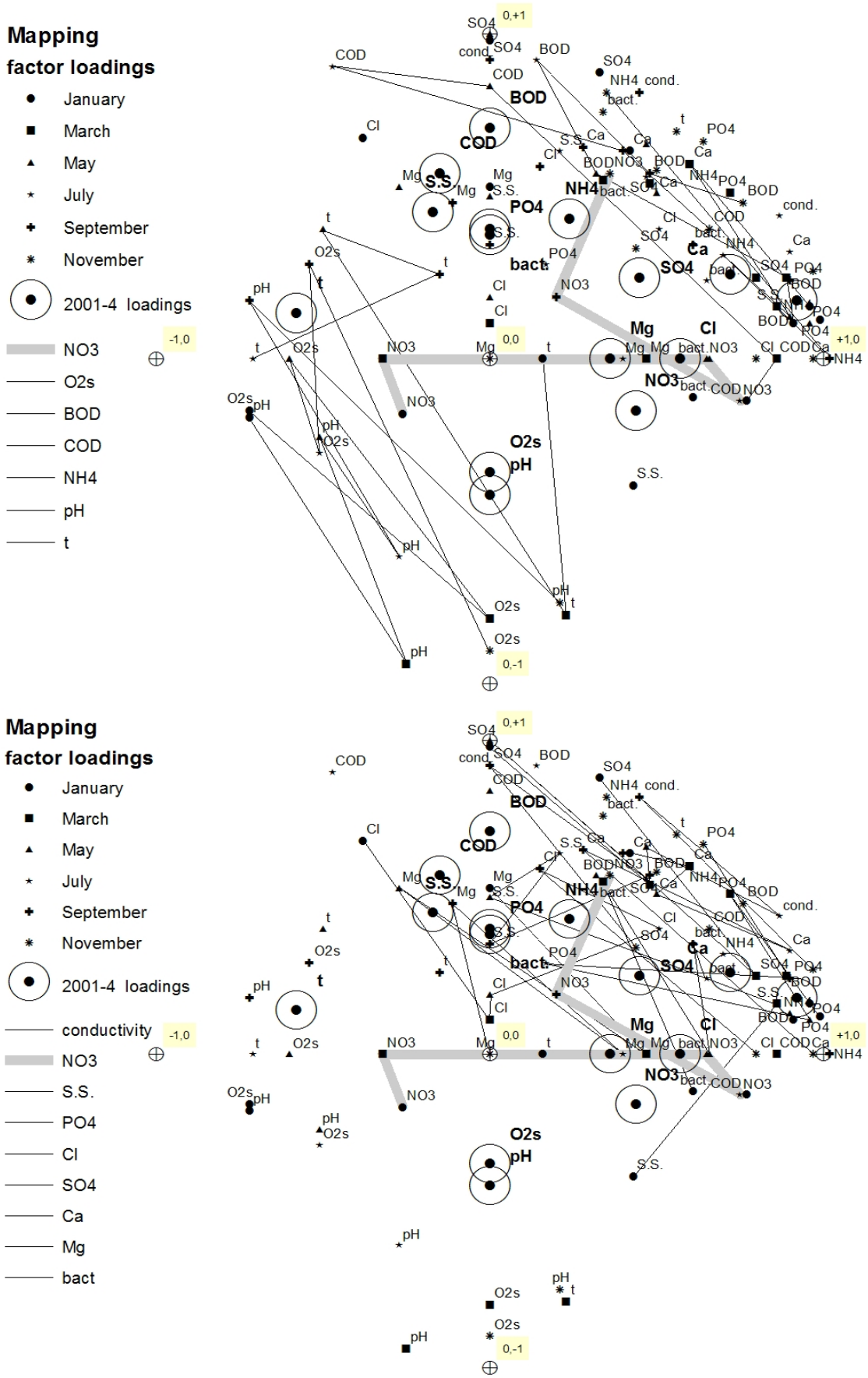


Figure 4. Examples of the mapping FA seasonal loading differences with the highlighted NO_3 path.