Information System for storing and processing data of environmental monitoring

Molorodov Y. I.[2], Minkov V.S.[12], Shirshov P.E.[12]

[1]Novosibirsk State University
Mechanics and Mathematics Department

[2]Institute of Computation Technologies SB RAS

ENVIROMIS, 2010

Promlems of time-series data exploration

Many aspects of human activity require researching of processes that dynamically change over time. Usually source data for these researches are time-series of some physical quantities: pressure, temperature, a substance concentration, etc.
Among the problems of ecological studies there is a class of subproblems, that needs such researchings, like:

- Monitoring of the atmosphere state of a large industrial center,
- Monitoring of multi-elemental composition of biosubstratums.

Information systems of ICT SB RAS

ICT SB RAS is solving some problems related to storing, processing and presentation time-series data of space distributed instumental observations.
Particularly such problems as:

- Creation of «Storing and researching system of cities and regions atmosphere state data»,
- Creation of «Siberian Biosubstratum» Atlas, intended for storing and processing data of multi-elemental blood composition of Siberians and inhabitants of the extreme north.

The are eleven air pollution observation posts in Novosibirsk at the moment, that are distributed in city and its suburbs and make regular sample probes of various atmosphere aerosols. As a result of these measurements (over 500 thousands recordings for each aerosol per year) time-series data transfers to Institute of Chemical Kinetics and Combustion SB RAS for further studies.

Also there are regular samplings of biosubstratums probes in various regions of Siberia, Khakassia, Buryatia, Far North, followed by measuring its multi-elemental composition by roentgen-fluorescent elemental analysis made on the station of elemental analysis in Centre of synchroton radiation of BINP SB RAS.

Source data

Source data of mentioned problems are time-series of scalar functions, that are associated with geographic coordinates of observation posts. These time series differs only by metadata sets however base metadata set is inherent for all the time-series:

- Coordinates of observation post,
- Measured quantity,
- Instrument, that was used for quantity measurement,
- Data preprocessing method.

Requirements

Following possibilities realization were needed within the context of projects:

- Importing of data, incoming from various organisations (ICK&C SB RAS, NIIC SB RAS, Novosibirsk Central Meteorological Service, BIC SB RAS) in huge number of different formats,
- Forming table reports for different time intervals and criterions,
- Visualisation of stored data and reports which are built using this data. (Observation posts representation on the map, different types of diagrams, etc),
- Stored data processing using various mathematically based algorithms. (cluster analysis, factor analysis, correlation analysis, wavelet analysis, etc)

Main problems and complications

- Nubmer of source formats using by data providers are growing,
- A need for new processing algorithms realization appears regulary,
- Source data amounts are rather big. For example, there are results of once-a-minute concentration measures of over ten aerosols for 2008-2009 and other years that are stored in the system(over 500 thousand recordings for each aerosol per year).

Principles and capabilities

A modular architecture was developed in order to solve a problem of regular expansion of system functionality need. Its main conception is an actively using of abstract interfaces, hooks and callbacks.
The core of the system grants a developer following capabilities:

- Module dependences and its solving,
- Register of modules, interfaces and realizations,
- Hooks and callbacks processing,
- User rights subsystem,
- Hierarchical menu generator,
- ORM adjustments for various SQL dialects,
- Deffered execution of resource-intensive tasks.

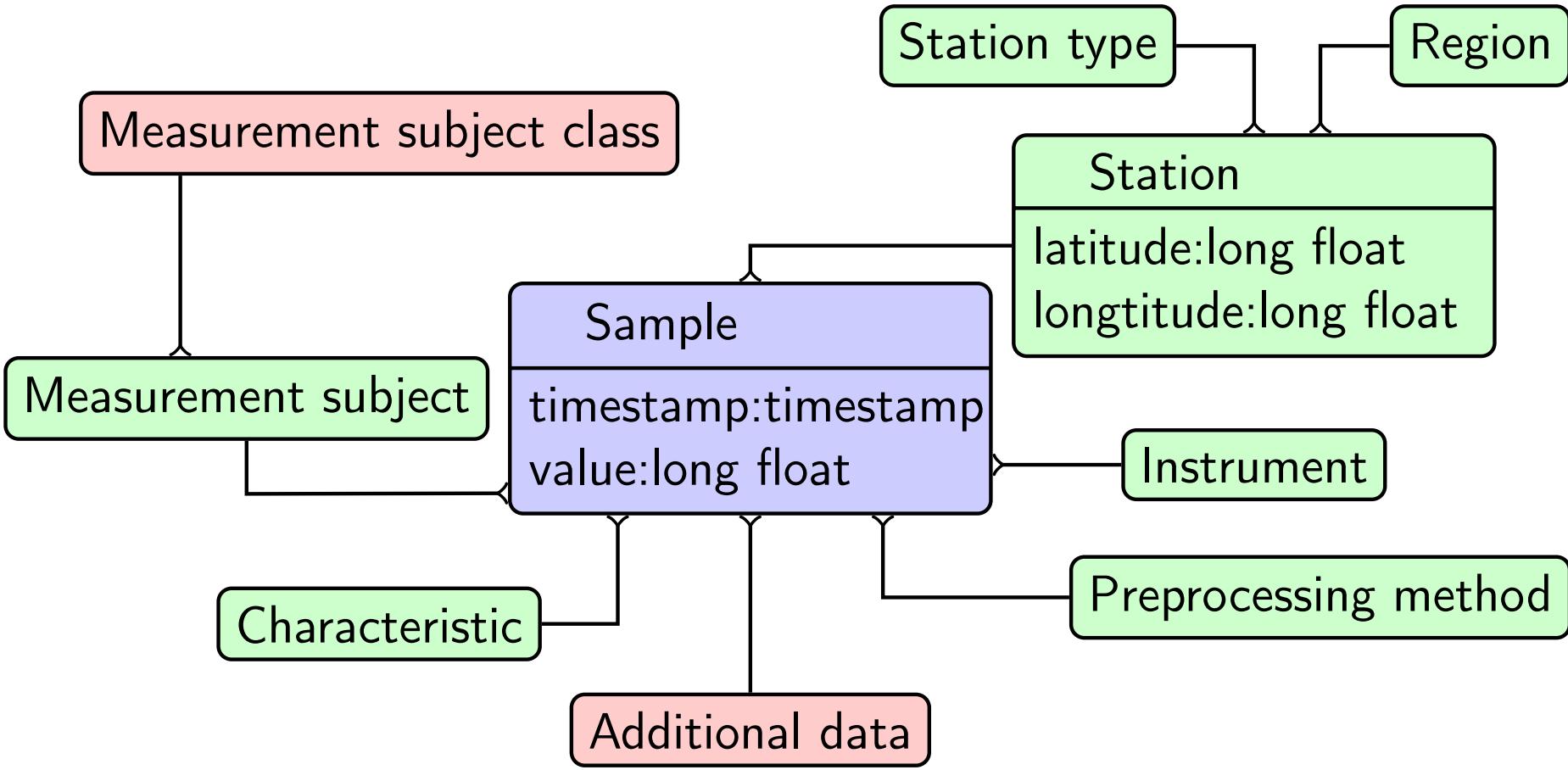
Extensibility

For a system functionality extending it is enough to create a new class, that implements one of the abstract interfaces provided by its basic modules.
For example new processing algorithm realization requires to implement one of two abstract interfaces provided by reports and processing subsystem. Implementations of the first interface can modify reports on its forming phase, implementations of second one can process already formed reports.
For changing a behaviour of already existing modules developer should create hooks and use callbacks.

Data model and its expansion

- Object-relational mapping (ORM) technology is used,
- The module of the basic data model provides an ability to work with its object representation,
- Time-series of quantities are representented as recordings of each single measures, its date and its basic set of metadata,
- Modules can extend the basic data model without any changes in its table structure. Information about extra metadata related to a measument is stored in separate tables. And it is dynamically associated with object representation of recordings with the use of ORM capabilities.

Logical representation of data model

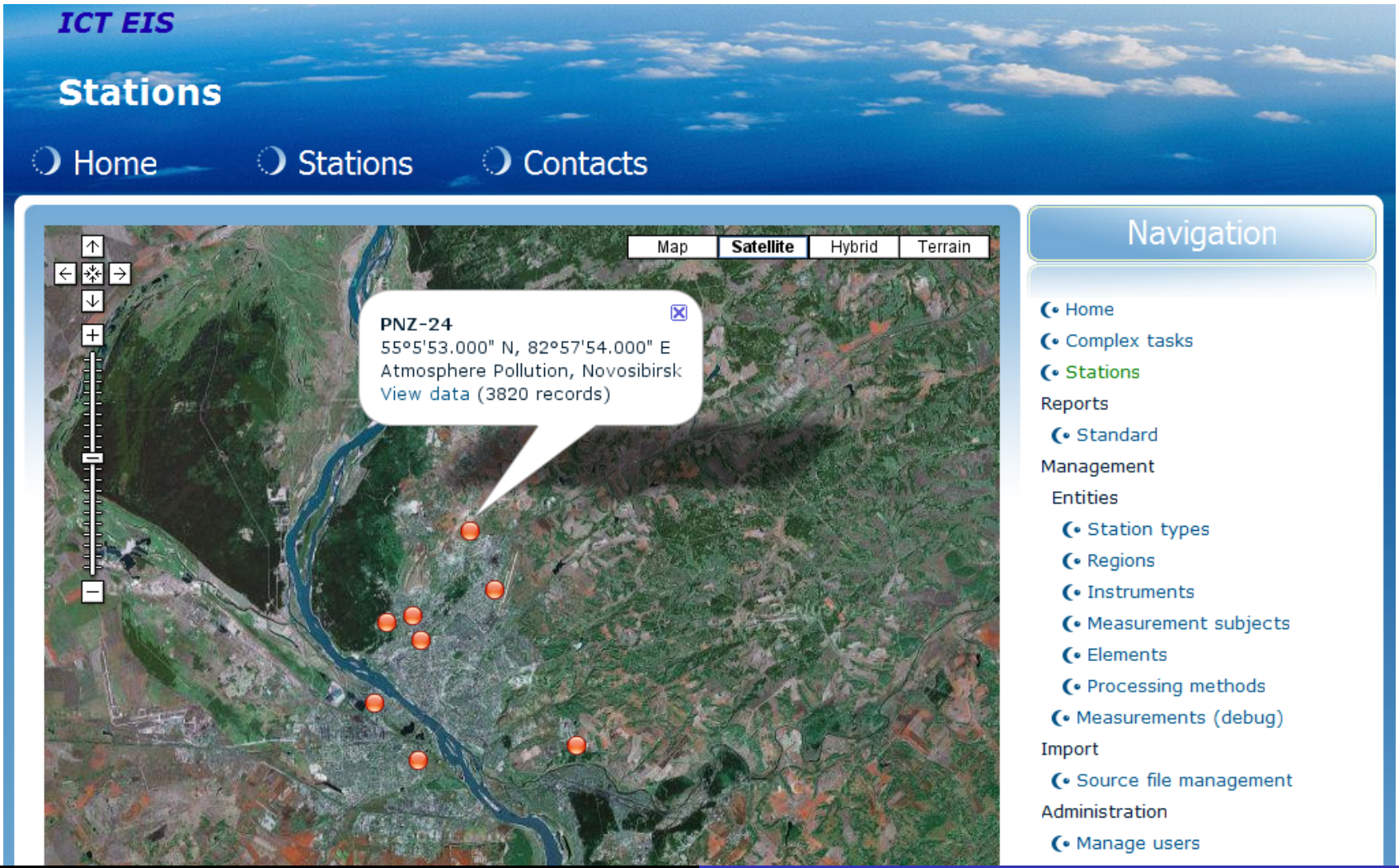


Cartographic module

Cartographic module allows user:

- To create, to delete and to edit information about observation posts,
- To view observation posts position on the map,
- To group observation posts by their types and to view only necessary types of posts,
- To proceed from station mark on the map to report generation for this station,
- To view properties of each observation post and number of measurements, associated with this post.

Station list



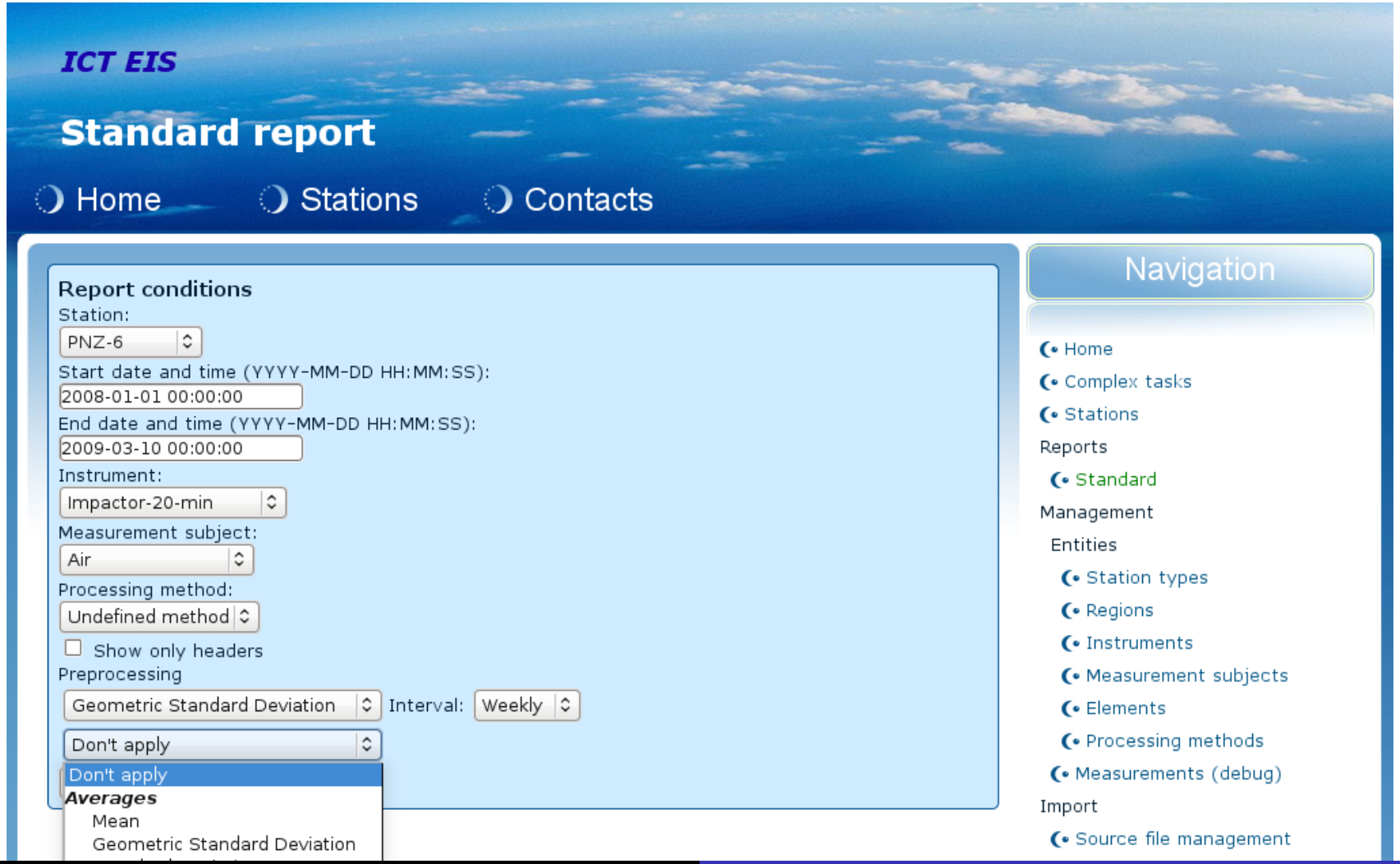
Subsystem of import

Basic subsystem of import provides control mechanisms for files uploaded on server and API for data entry in DB.
Implementations of abstract interface provided by subsystem give a support for specific formats.

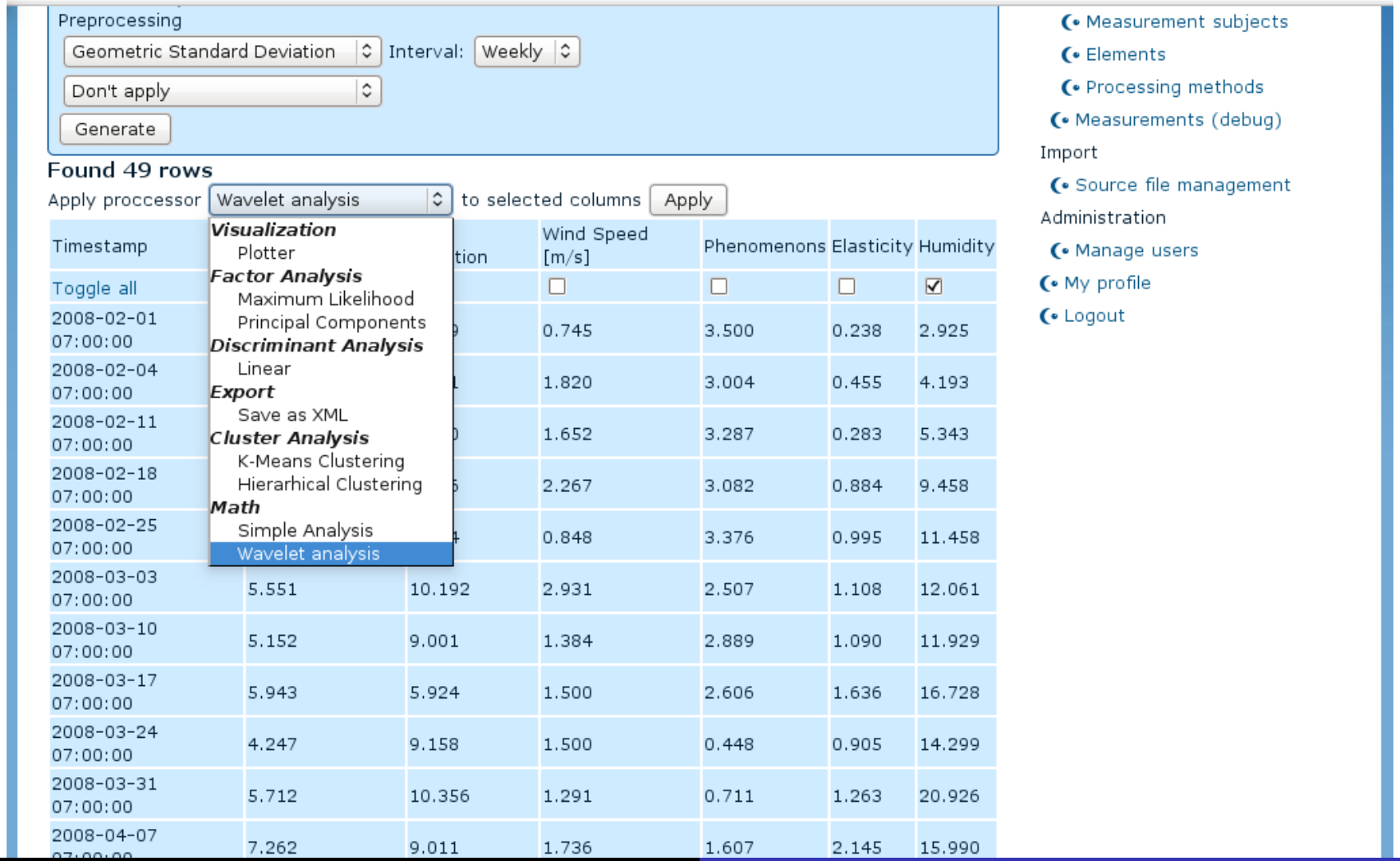
Subsystem of reports and processing

Allows to create table reports by various criterions in which all measurement metadata can take part. Can be extended by implementation of two abstract interfaces realization — *preprocessor* and *postprocessor*.
Preprocessors are used for modifying a report on its forming phase (for example, averaging by various periods or missing values approximation).
Postprocessor are used for already formed reports processing.
User interface allows to chose arbitrary set of preprocessors and postprocessors, that will be applied to report. Each preprocessor and postprocessor can modify a form of report creation for missing parameters requesting.

Process of report formation



Processing of a complete report



Powered with

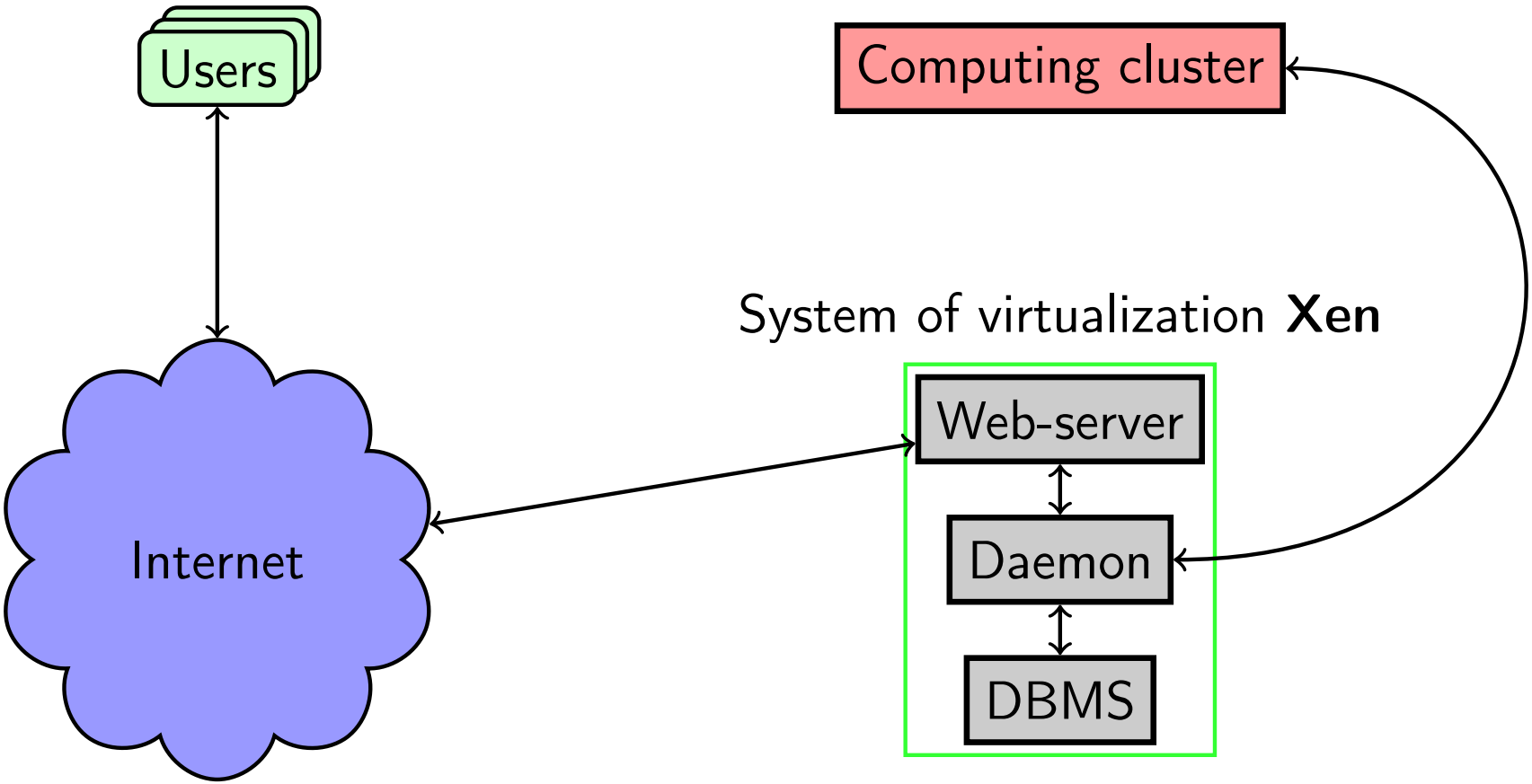
The system is powered with:

- **Python** programming language,
- **Pylons** web-framework,
- **SQLAlchemy** object-relational mapper (ORM),
- **XHTML 1.0 Transitional** and **jQuery** library,
- **C++** programming language and **OpenMP** technology,
- **Google API** for representation of geographical information.

The following software was used as a sources of data processing algorithms:

- **R** programming language,
- **Numpy** and **Scipy** libraries,
- Own products of ICT SB RAS.

Basic structure



Software

Main server of the system, SQL server and web-server are virtualized by hypervisor. **Xen 3.4.2**, **Gentoo Linux** is used as a OS for host and guest nodes.

Following software are used for operation of system version for ultimate users:

- Python: **CPython 2.6**,
- OS: **Gentoo Linux**,
- DBMS: **PostgreSQL 8.4**,
- Web-server: **nginx** as a frontend, **cherrypy** as a WSGI-backend.

Queue for resource-intensive tasks for its using on cluster was realised by: **GNU Screen**, **Bash** and **OpenSSH**.

Hardware

Hypervisor that provides operationality of the system's virtual machines use server with following characteristics:

$$4 \times \text{Intel Xeon @ 2.8 GHz, 3 Gb RAM}$$

Computing cluster **MIST** based on **Tyan VX50** platform, that is used for resource-intensive tasks' execution, is located in ICT SB RAS and has following characteristics:

$$8 \times \text{Dual Core AMD Opteron @ 2.5 GHz, 32 Gb RAM}$$

Wavelet-analysis (Morlet wavelet)

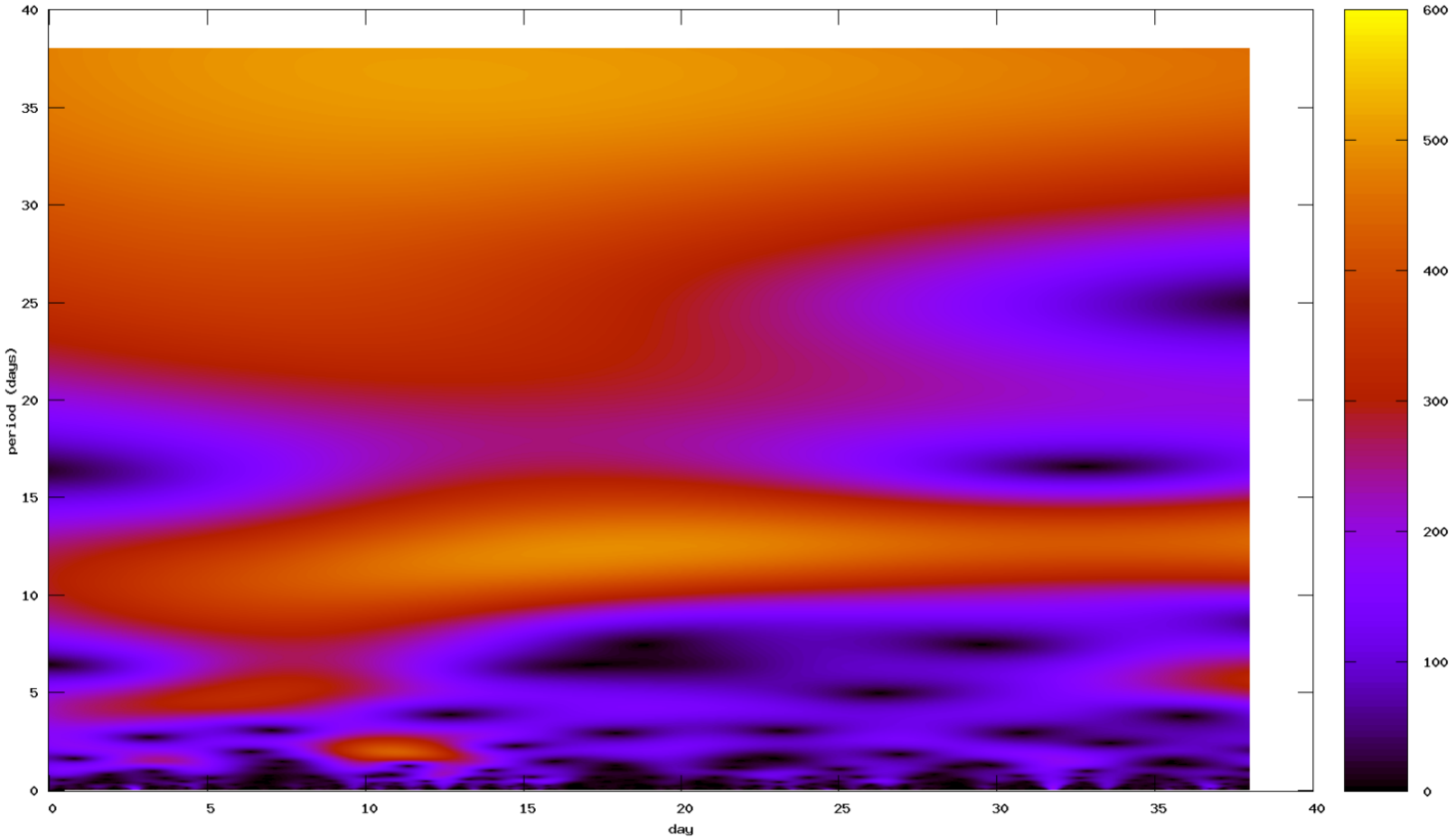
$$W(t,p) = \frac{1}{p}\int_{-\infty}^{+\infty} \overline{\psi}\left(\frac{x-t}{p}\right) f(x)\,dx$$

$$\psi(\theta) = \pi^{-\frac{1}{4}} e^{-i\omega_0\theta} e^{-\frac{\theta^2}{2}}$$

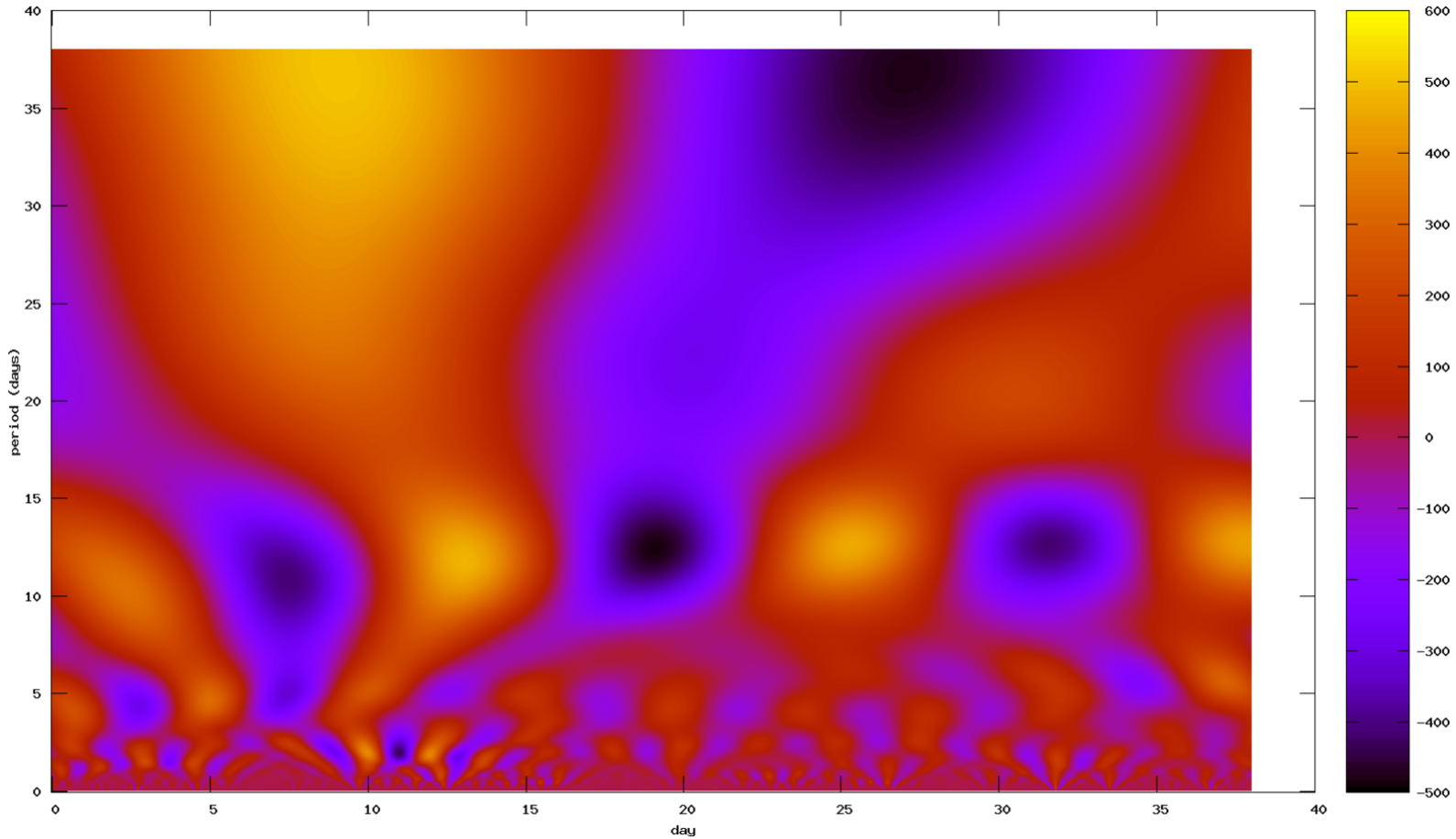
Data set description

- **Station**: Klyuchi;
- **Coordinates**: 54°46´31”N, 83°5´52”E;
- **Analyzed quantity**: submicron fraction atmospheric aerosol, $\frac{mcg}{m^3}$;
- **Time range**: 2009-01-01 – 2009-02-07;
- **Number of samples**: 54720 samples during 38 days

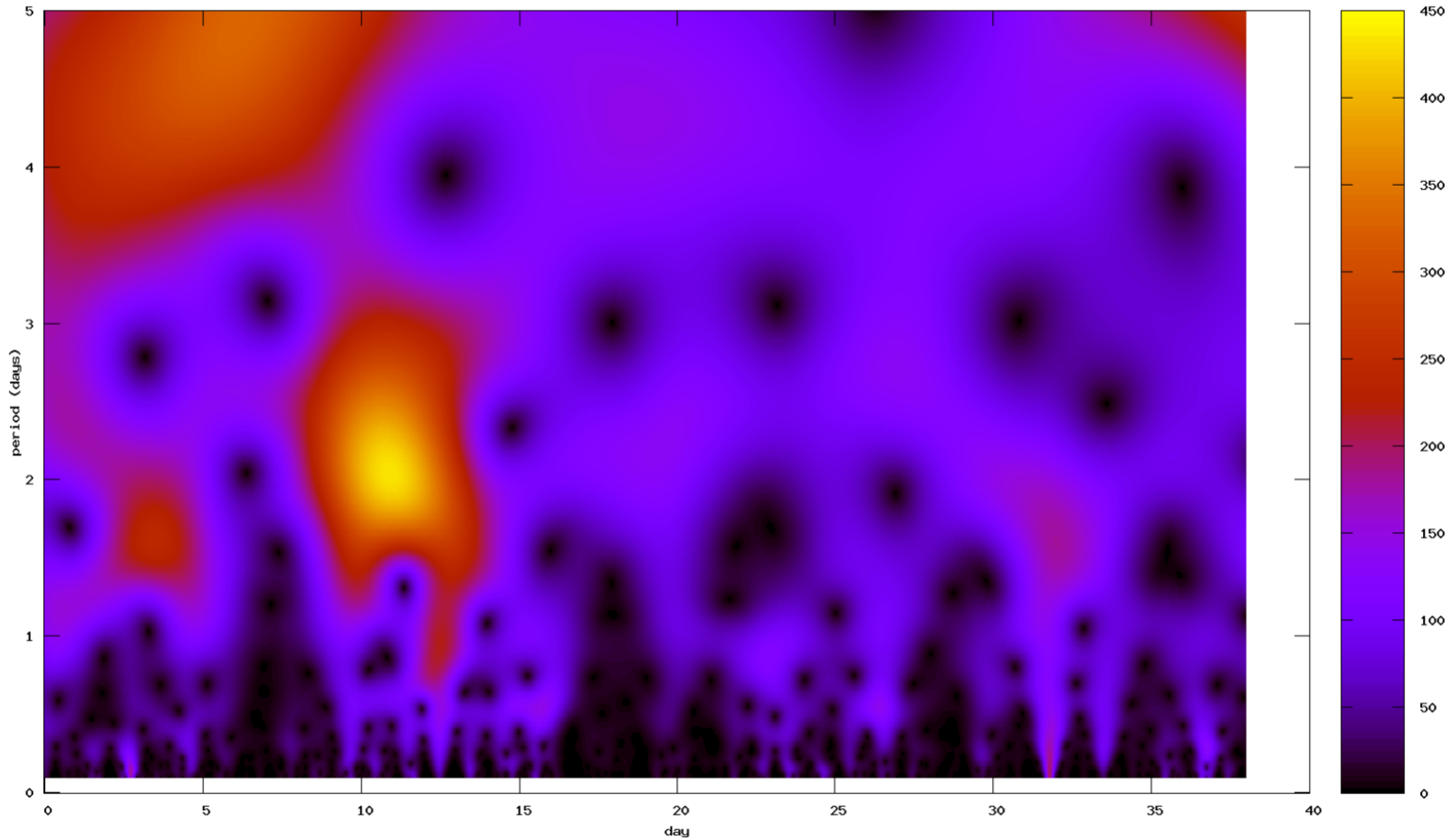
Winter 2009. Modulus of morlet wavelet



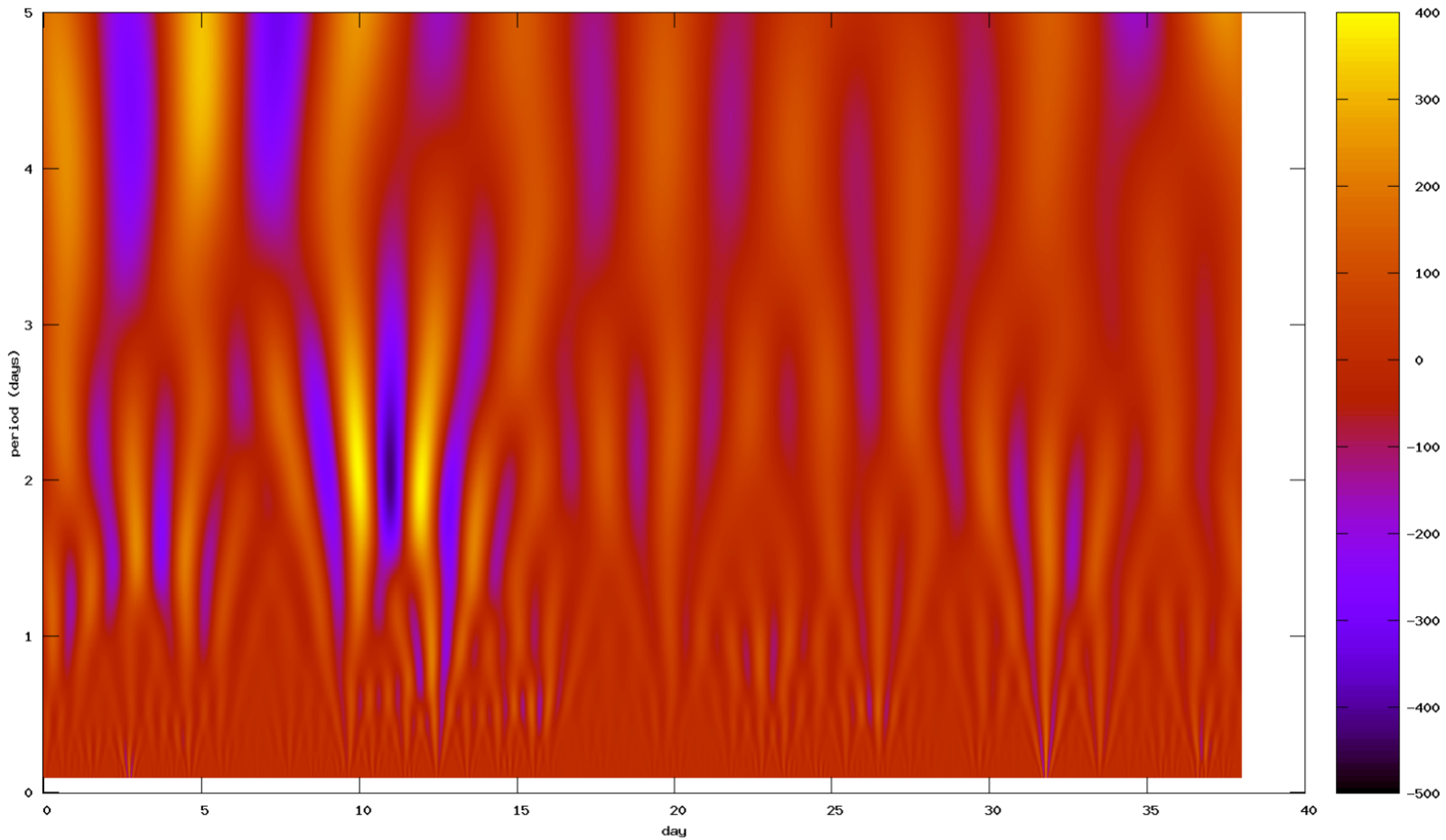
Winter 2009. Real part of morlet wavelet



Winter 2009. Modulus of morlet wavelet. Short periods



Winter 2009. Real part of morlet wavelet. Short periods

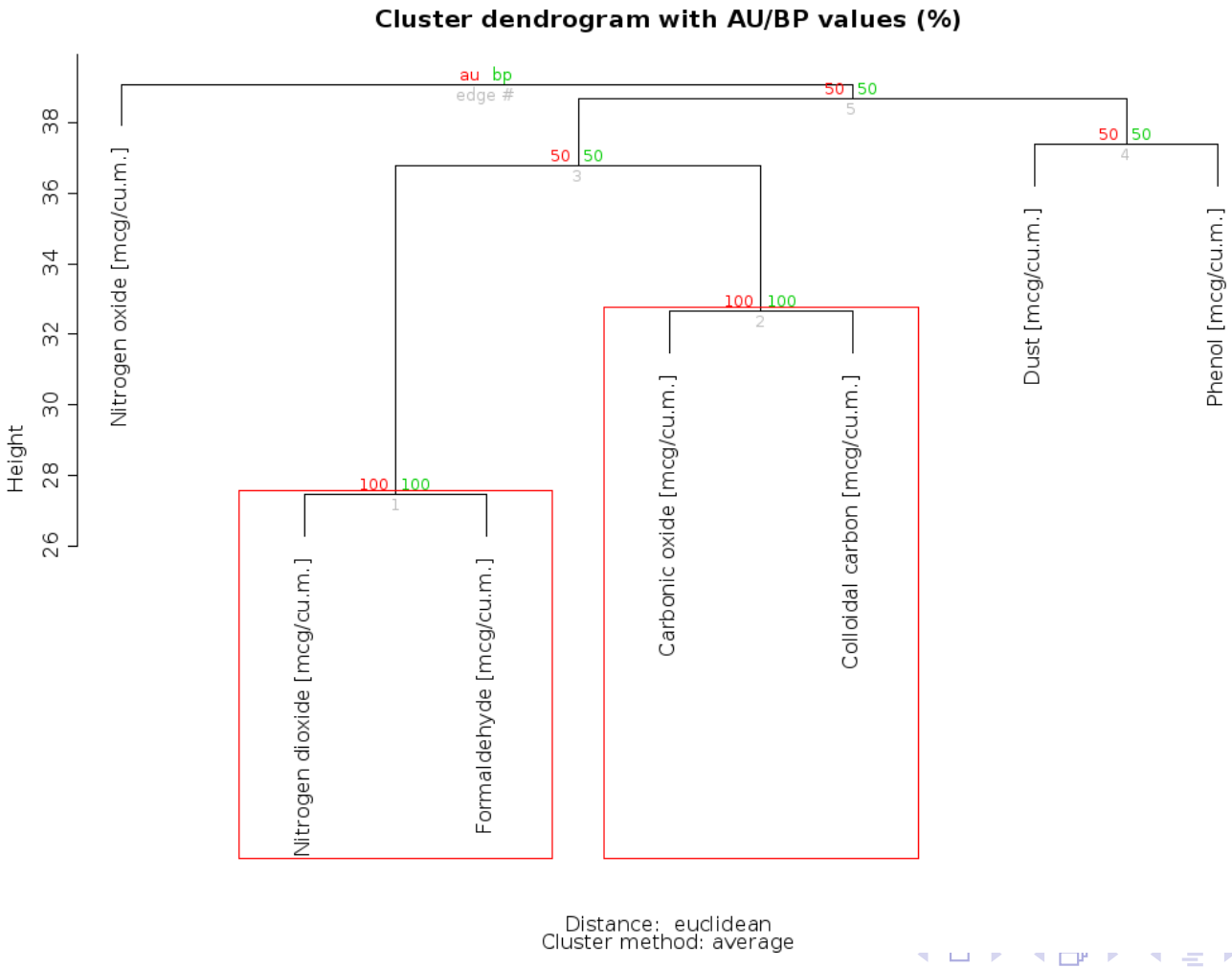


Hierarchical clusterisation with cross-validation algorithm (multiscale bootstrap resampling)

Data set description

- **Station**: PNZ-25;
- **Coordinates**: 54°57´56.000"N, 82°54´45.000"E
- **Analyzed quantities**:
  Dust, $\frac{mcg}{m^3}$
  Carbonic oxide, $\frac{mcg}{m^3}$
  Nitrogen dioxide, $\frac{mcg}{m^3}$
  Nitrogen oxide, $\frac{mcg}{m^3}$
  Phenol, $\frac{mcg}{m^3}$
  Colloidal carbon, $\frac{mcg}{m^3}$
  Formaldehyde, $\frac{mcg}{m^3}$;
- **Instrument**: impactor;
- **Time range**: 2008-01-26 – 2008-12-31;
- **Number of probes**: 839.

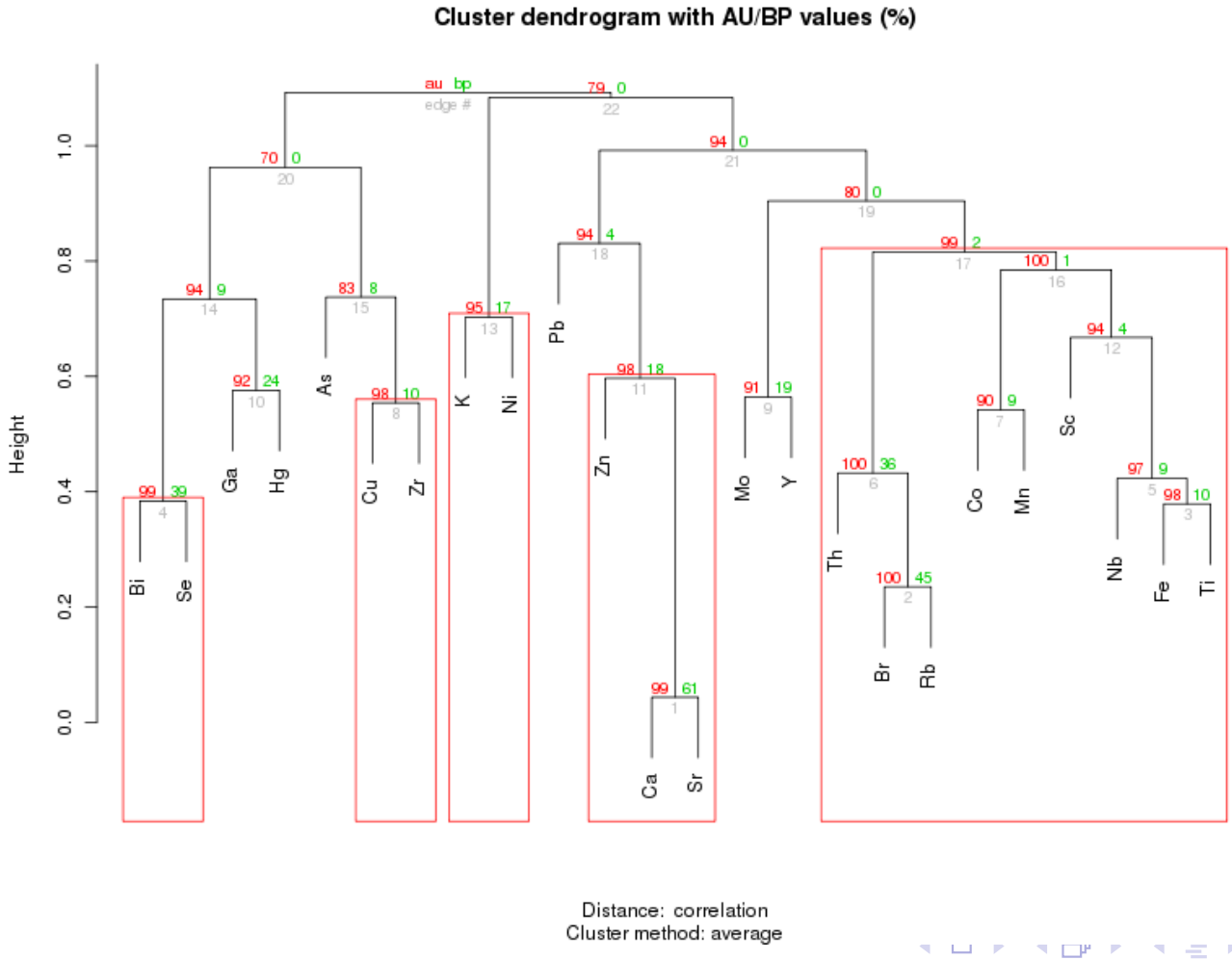
Result of the processing



Data set description

- **Station**: Krasnoselkup village (mobile researching station;);
- **Analyzed quantities**:
  multi-elemental composition of Nenets representatives' blood, $\frac{mcg}{m^3}$
- **Instrument**: Synchrotron Radiation X-Ray Fluorescence Analysis (SRXRF);
- **Time range**: 2007-04-05 – 2007-07-03;
- **Number of probes**: 126.

Result of the processing



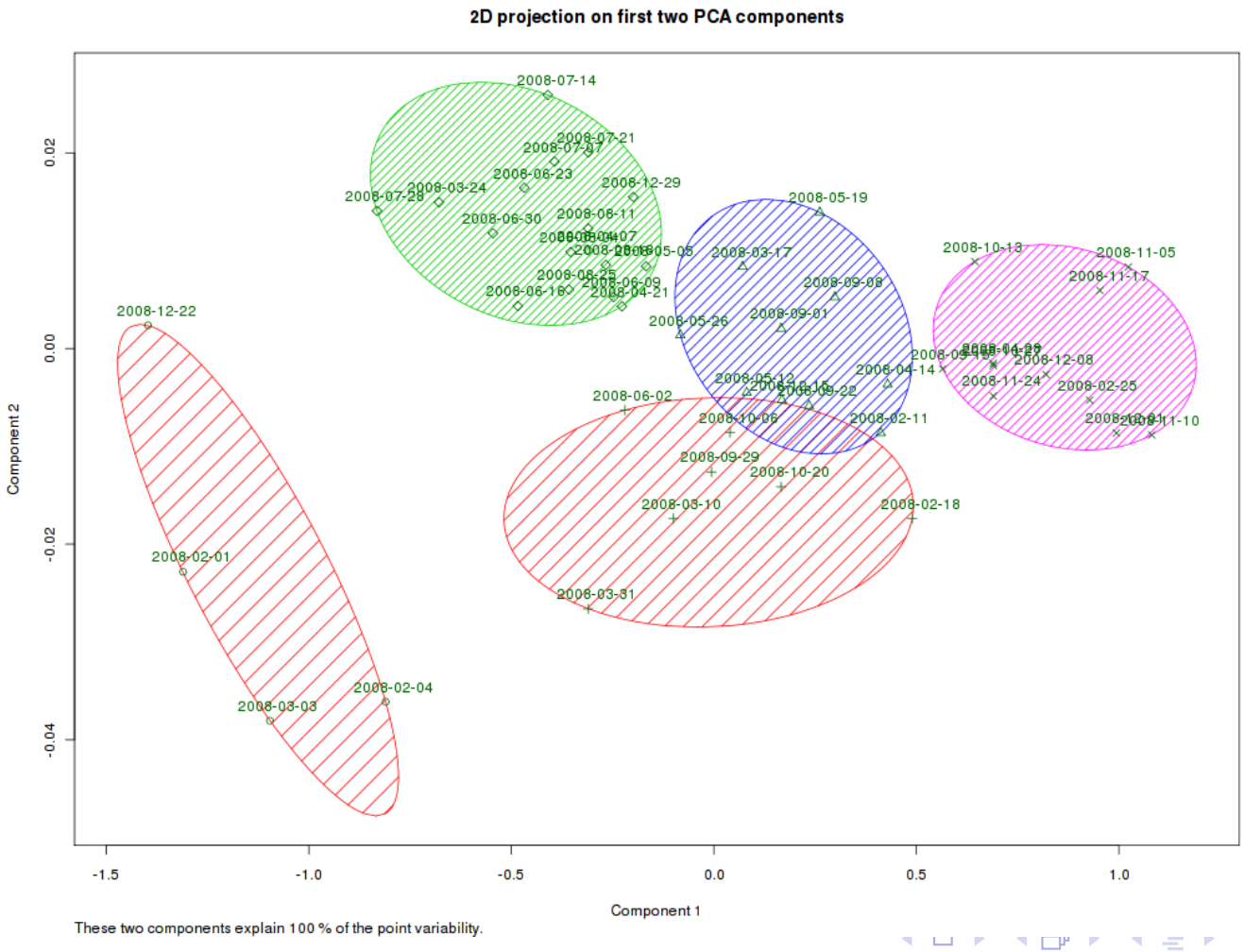
k-means clustering (Hartigan-Wong, KMeans++ algorithm)

$$\phi = \sum_{i=1}^{K} \sum_{j=1}^{p} \sum_{m=1}^{n_i} f_{\nu_{im}} \omega_{\nu_{im}} \delta_{\nu_{im},j} (x_{\nu_{im},j} - \bar{x}_{ij})^2$$

Data set description

- **Station**: PNZ-26;
- **Coordinates**: 55°2′56.000″N, 82°54′23.000″E;
- **Analyzed quantity**: measurement time-series;
- **Instrument**: impactor;
- **Time range**: 2008-01-09 – 2008-12-31;
- **Number of probes**: 1092.

Result of the processing



These two components explain 100 % of the point variability.

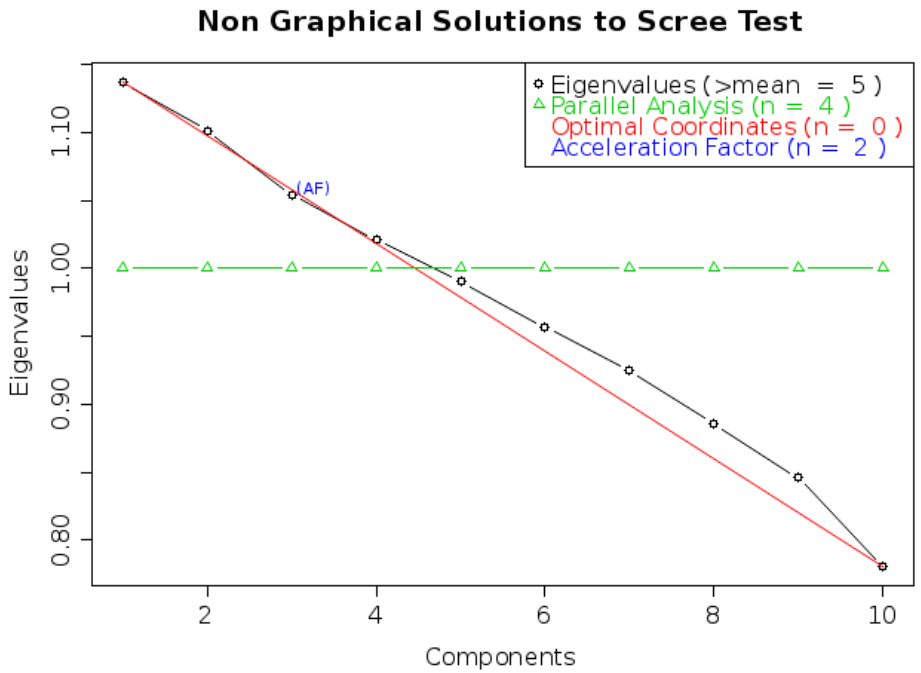
Principal Component Analysis

$$\mathcal{L}(\theta \mid x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i|\theta)$$

Data set description

- **Station**: PNZ-21;
- **Coordinates**: 55°2′43.000″N, 83°52′53.000″E;
- **Analyzed quantities**:
  Dust, $\frac{mcg}{m^3}$, Carbonic oxide, $\frac{mcg}{m^3}$, Nitrogen dioxide, $\frac{mcg}{m^3}$, Colloidal carbon, $\frac{mcg}{m^3}$, Ammonia, $\frac{mcg}{m^3}$, Formaldehyde, $\frac{mcg}{m^3}$, Temperature, $^{\circ}C$, Wind Direction, Wind Speed $\frac{m}{s}$, Phenomenons;
- **Instrument**: impactor;
- **Time range**: 2008-01-09 – 2008-12-31;
- **Number of probes**: 1087.

Result of the processing



Additional info

The number of observations is 768

Test of the hypothesis that 2 factors are sufficient.
The chi square statistic is 192.656 on 26 degrees of freedom.
The p-value is 0.000

Importance of components

	F1	F2
SS Loadings	1.23	1.10
Proportion of Variance	0.12	0.11
Cumulative Proportion	0.12	0.23

	F1	F2	Uniqueness
Dust [mcg/cu.m.]	0.20	0.10	0.950
Carbonic oxide [mcg/cu.m.]	*-0.03*	0.26	0.932
Nitrogen dioxide [mcg/cu.m.]	*-0.06*	0.61	0.624
Colloidal carbon [mcg/cu.m.]	*-0.38*	0.14	0.837
Ammonia [mcg/cu.m.]	0.16	*-0.09*	0.968
Formaldehyde [mcg/cu.m.]	*0.02*	**0.75**	0.442
Temperature [°C]	**0.97**	0.24	0.005
Wind Direction	*0.05*	*-0.04*	0.996
Wind Speed [m/s]	*-0.05*	*-0.08*	0.992
Phenomenons	*-0.27*	*-0.05*	0.927

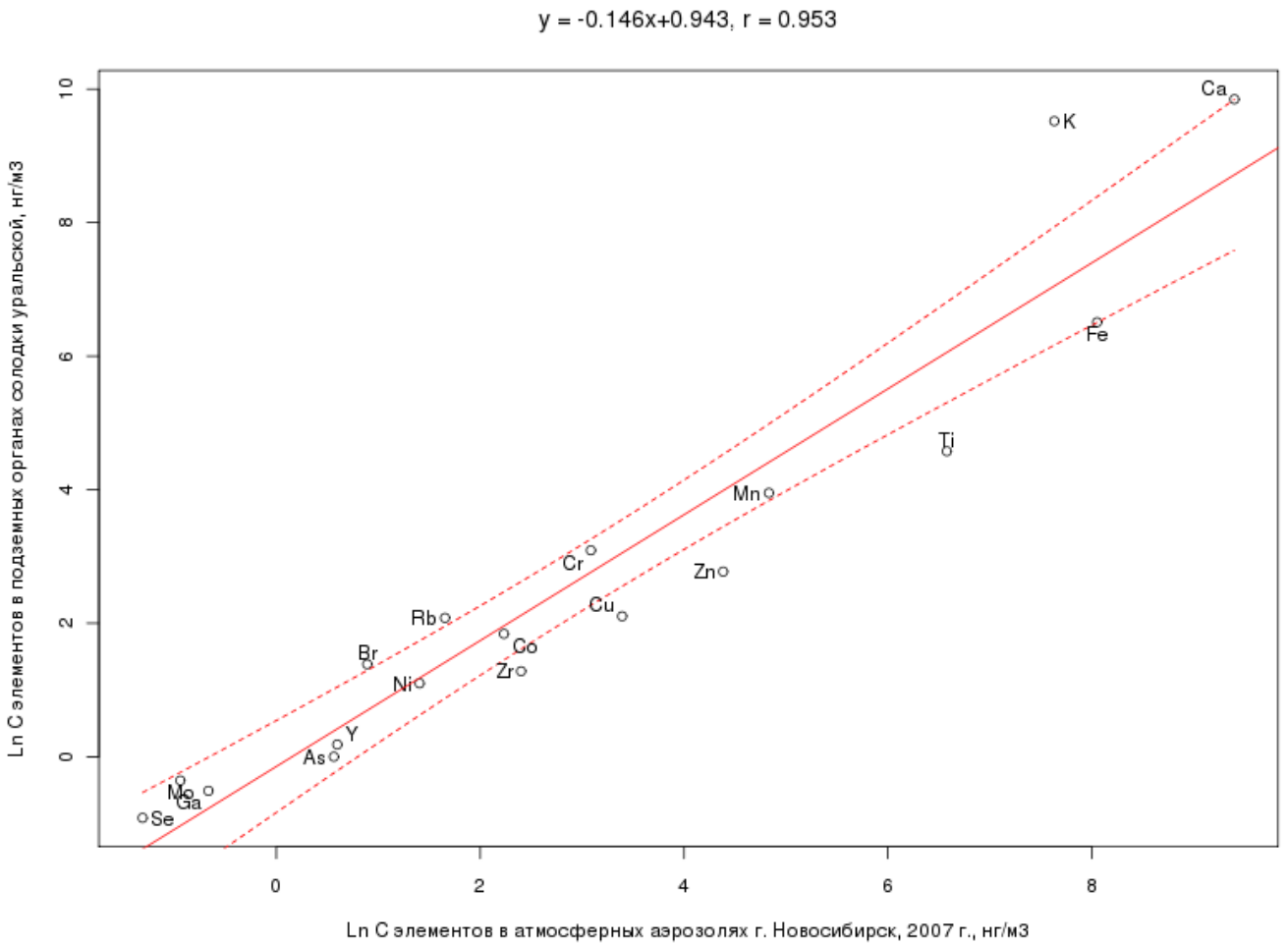
Linear regression

$$\mathcal{L}(\theta \mid x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i|\theta)$$

Data set description

- **Station**: mobile researching station;
- **Analyzed quantities**: multi-elemental composition of licorice (*Glycyrrhiza uralensis*) and air, Novosibirsk;
- **Instrument**: Synchrotron Radiation X-Ray Fluorescence Analysis (SRXRF);
- **Time range**: 2007-05-15 – 2007-07-02;
- **Number of probes**: 115.

Result of the processing

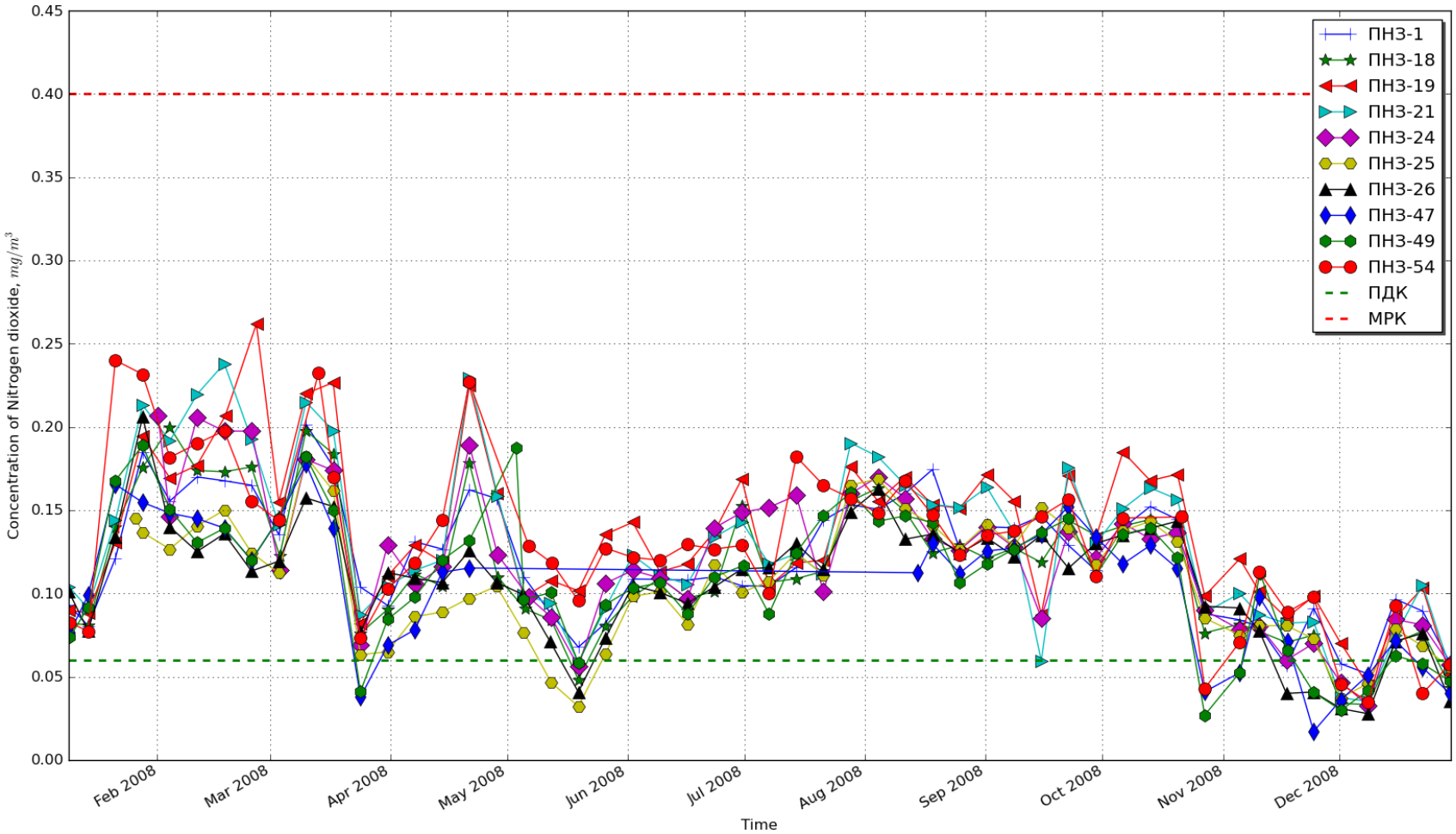


Visualisation with plots

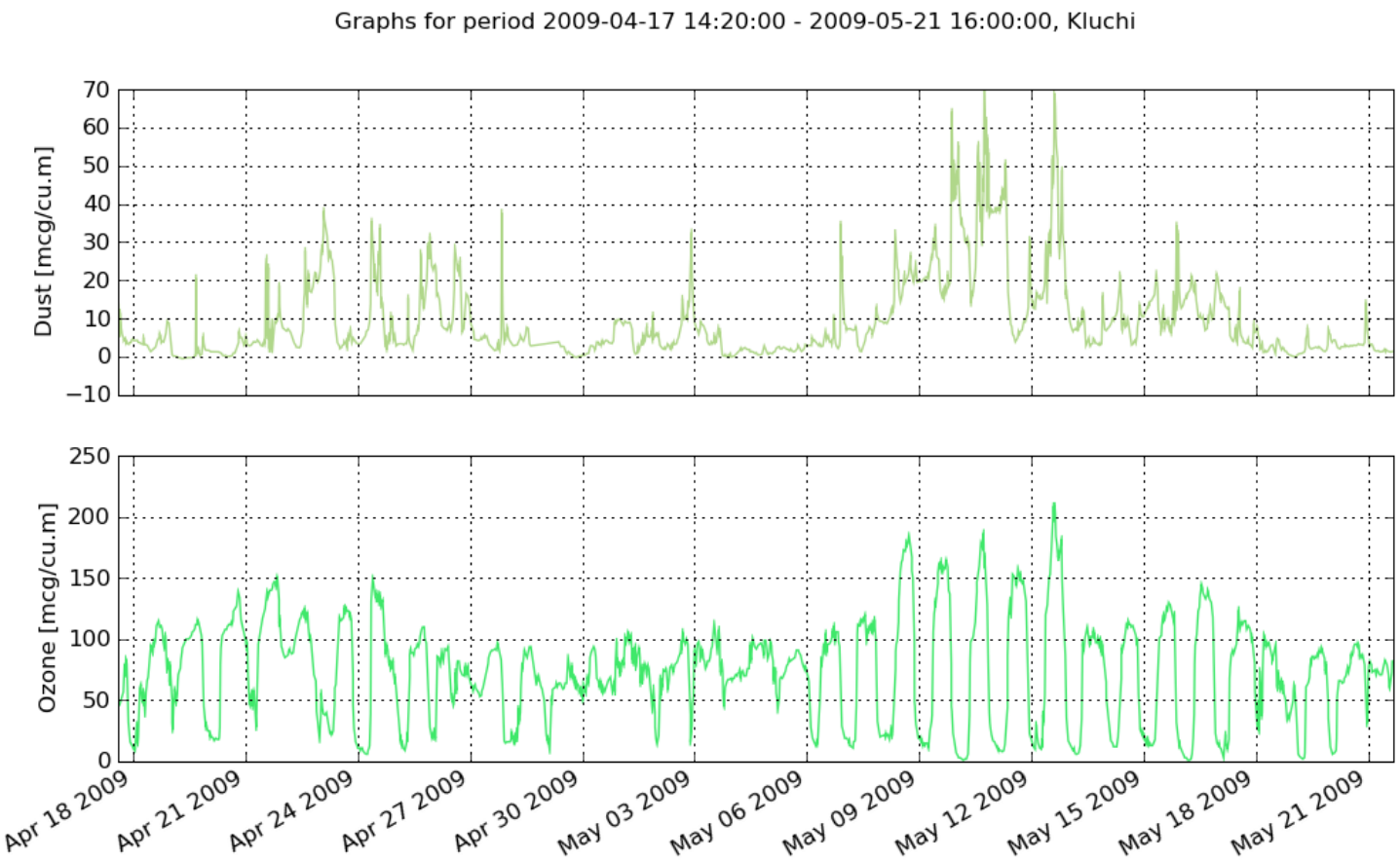
Data set description

- **Analyzed quantities**:
  NO2, $\frac{mcg}{m^3}$;
- **Time range**: 2008-01-01 – 2008-12-31;
- Time series from all the stations was used.

NO2 concentrations



Dust/Ozone concentrations



Concluding part

Within the bounds of project

- Modular platform for building of web-oriented informational systems was created,
- Extendible data model for storage of scalar time-series was developed,
- Extendible system for storing and forming already stored reports was created,
- Capabilities for mathemathical processing of stored data and for other system components creation were provided.

Thank you for your attention!