Exploratory Data Analysis for Ecological Modelling and Decision Support

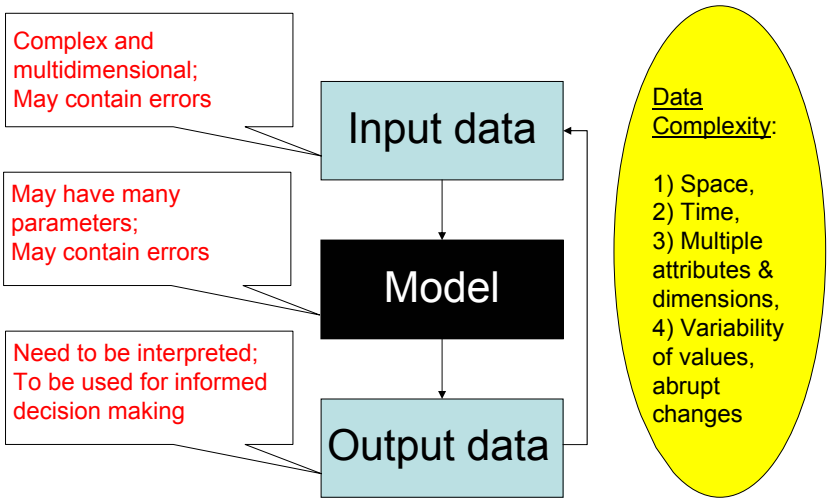
Gennady Andrienko & Natalia Andrienko
Fraunhofer Institute AIS
Sankt Augustin
Germany
http://www.ais.fraunhofer.de/and

Fraunhofer Institut Autonome Intelligente Systeme

Outline

1. Geo-visualisation's view on ecological modelling: demanding problems and challenging tasks
2. Case study 1: pesticide accumulation
3. Case study 2: forest dynamics
4. A systematic approach to exploratory data analysis (EDA): elements of the general theory
5. Software issues

A View on Ecological Modelling



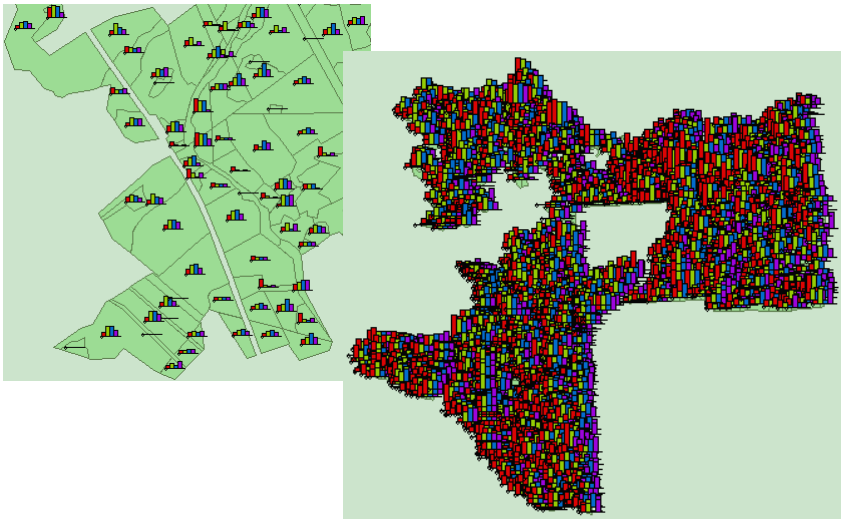
Outputs of simulations

- Multiple attributes referring to
  - Simulation scenarios;
  - Spatial locations (objects);
  - Time moments;
  - … (e.g. species, age groups etc.)

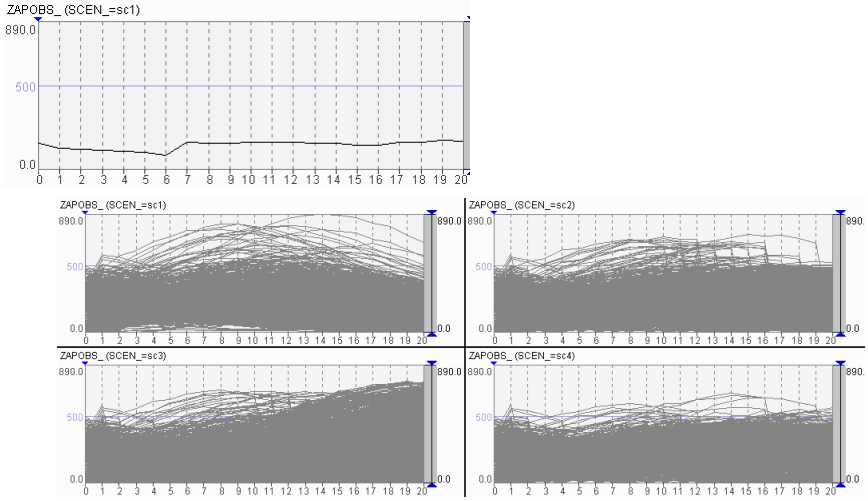
Complexities

- Number of attributes
- Length of time series
- Number of spatial objects
- High dimensionality => huge number of combinations (normally 10^5-10^8)!
- Abrupt temporal changes
- Great variability of values

Complexities: example 1



Complexities: example 2



Our goals

Support analysts and decision makers in:

- Preparing and harmonizing input data;
- Tuning models and their parameters;
- Interpreting outputs of simulation;
- Exploring alternatives for decision making;
- Justifying and communicating the resulting decisions.

Instruments: interactive visualisation enhanced by intelligent aggregation tools and other tools for exploratory data analysis

Outline

1. Geo-visualisation's view on ecological modelling: demanding problems and challenging tasks
2. Case study 1: pesticide accumulation
3. Case study 2: forest dynamics
4. A systematic approach to exploratory data analysis (EDA): elements of the general theory
5. Software issues

GIMMI project

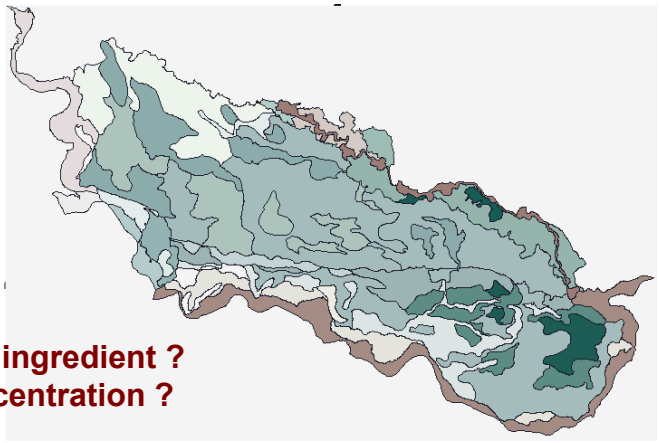
Geographical Information and Mathematic Model Interoperability

IST-2001-34245, 2002 – 2004

TXT Italy, EIG Germany, AIS Germany…

- ✓ Multiple simulation scenarios (different crops and active ingredients)
- ✓ about 1000 plots
- ✓ simulation depth: 10+ years
- ✓ several output variables that characterize various environmental aspects (pesticide accumulation etc.)

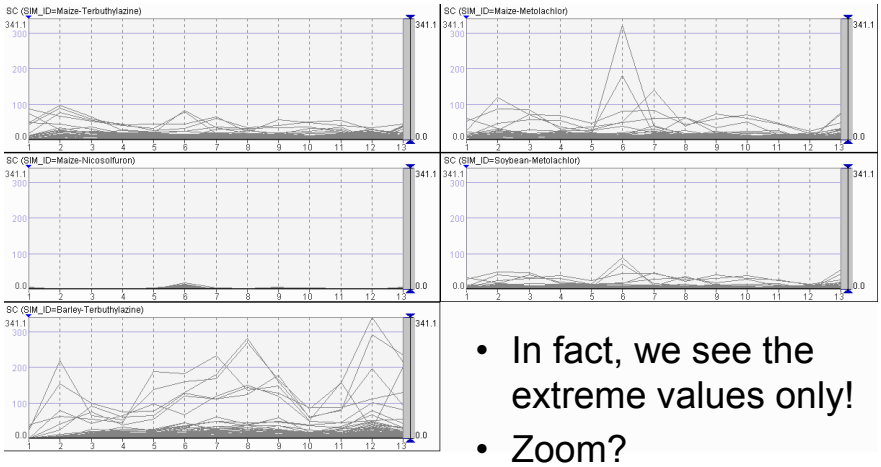
Decisions to be made



- ✓ **What crop ?**
- ✓ **What active ingredient ?**
- ✓ **In what concentration ?**

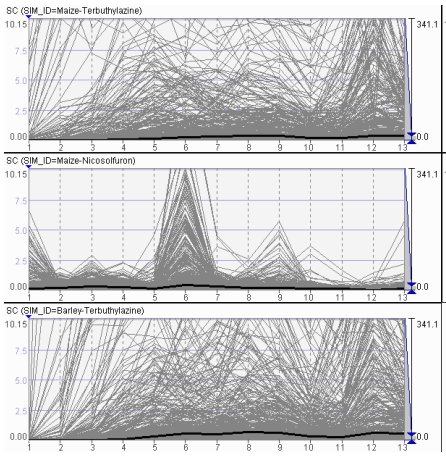
- ➢ **… for individual plots**
- ➢ **… and for the whole territory**

Pesticide accumulation dynamics



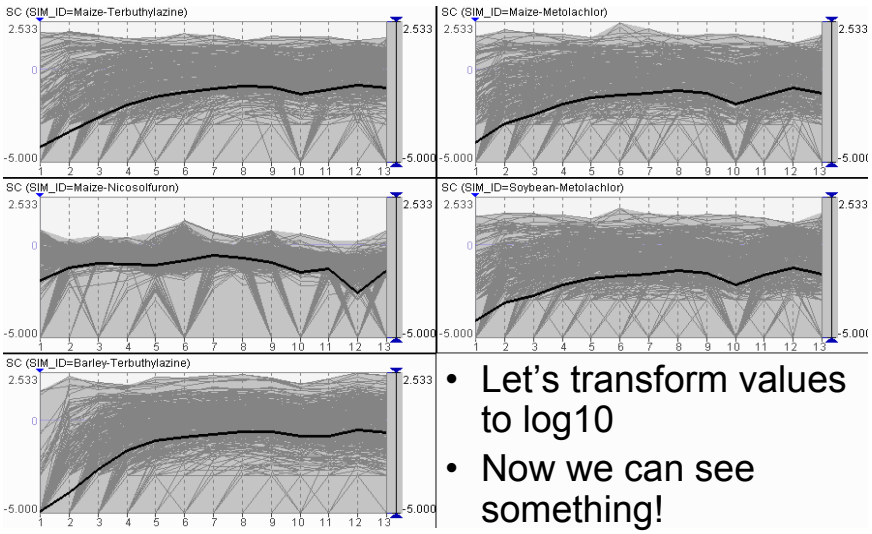
- In fact, we see the extreme values only!
- Zoom?

Zoomed pesticide accumulation



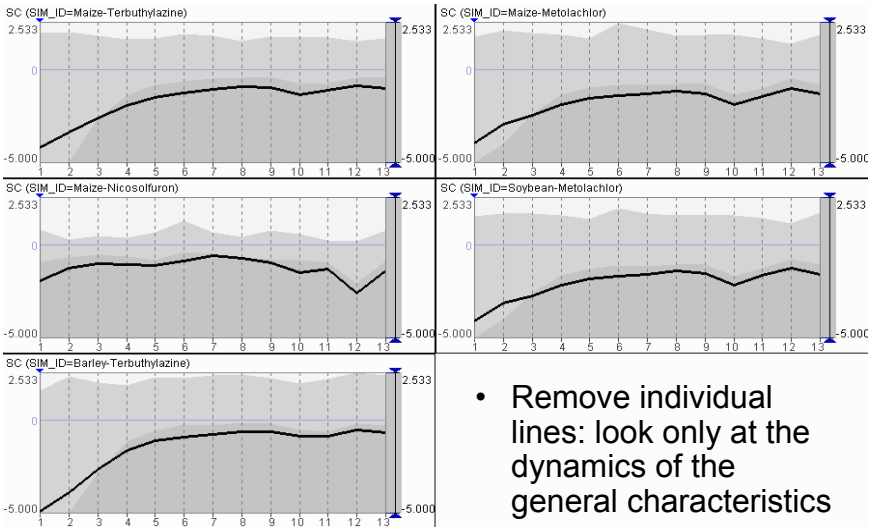
- The extreme values and the overall view are lost
- But the details are still not visible!

Log10 transformed values



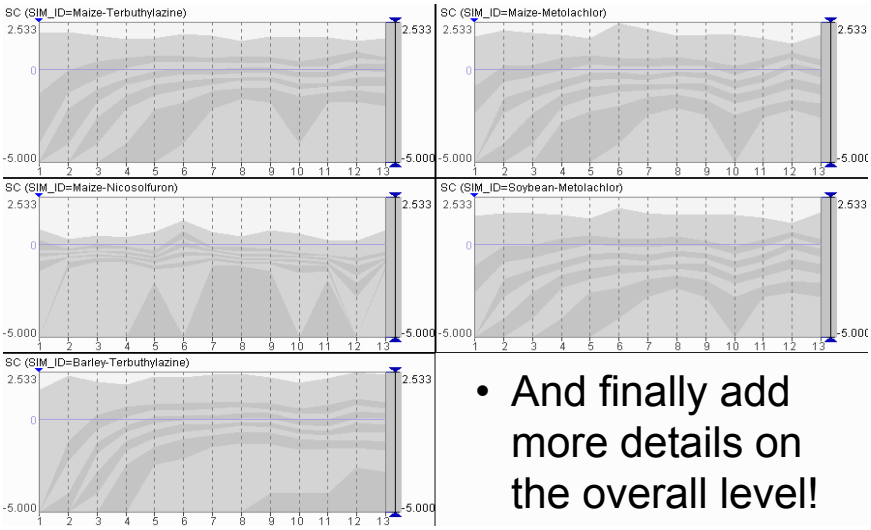
- Let's transform values to log10
- Now we can see something!

Only means, medians, and envelopes



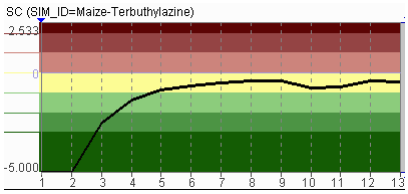
- Remove individual lines: look only at the dynamics of the general characteristics

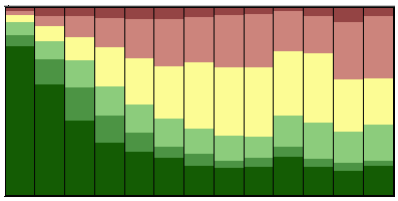
Details of the distribution



- And finally add more details on the overall level!

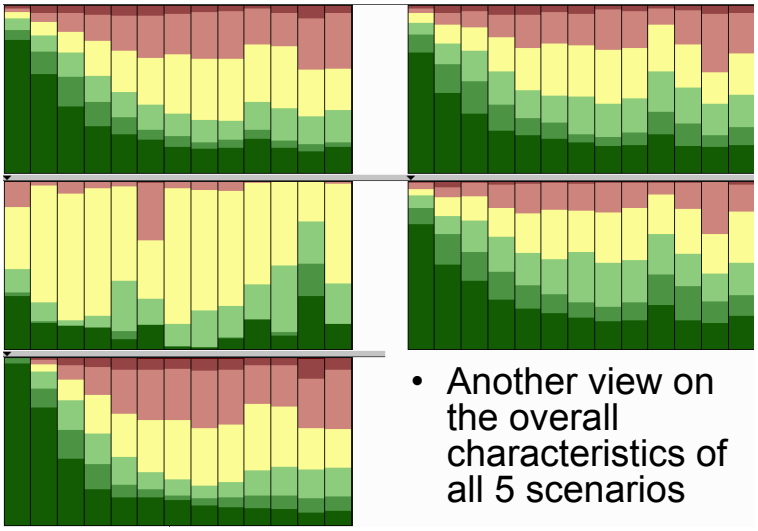
Count plots within intervals





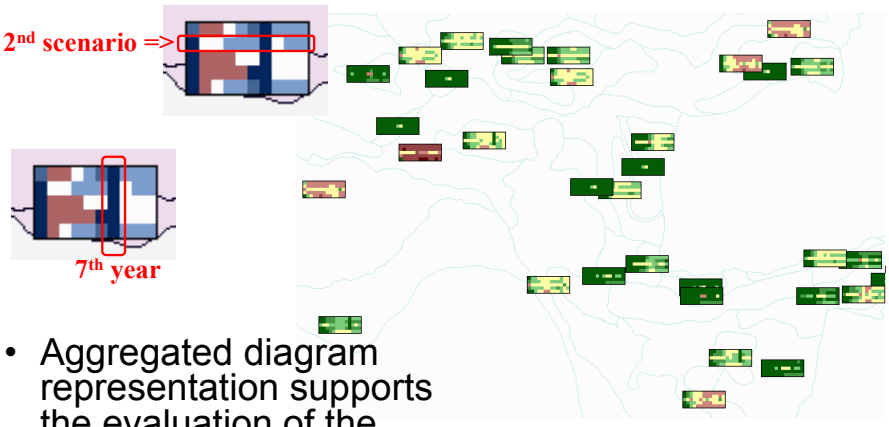
1. Specify classes according to pesticide concentration
2. Count number of plots within each interval for each year
3. Draw the counts as stacked bars
4. Possible extension: use areas or other amounts instead of the counts

Compare the accumulation in all scenarios for the whole territory



- Another view on the overall characteristics of all 5 scenarios

Look at the individual plots



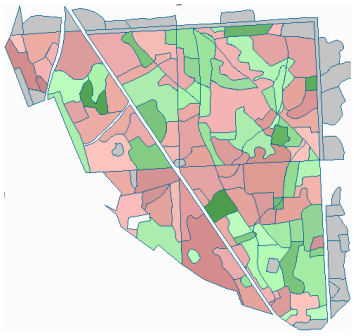
- Aggregated diagram representation supports the evaluation of the scenarios for individual plots

Dynamic linking between displays supports selection of "interesting" plots

Outline

1. Geo-visualisation's view on ecological modelling: demanding problems and challenging tasks
2. Case study 1: pesticide accumulation
3. Case study 2: forest dynamics
4. A systematic approach to exploratory data analysis (EDA): elements of the general theory
5. Software issues

Silvics project

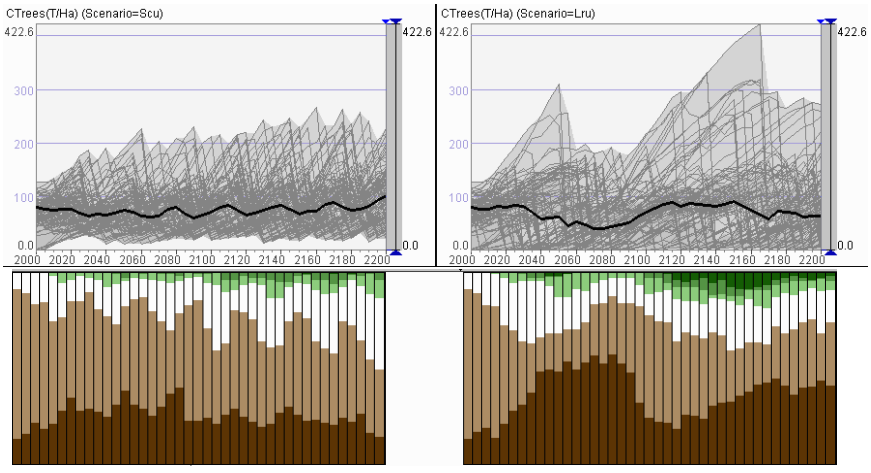


- ✓ 104 forest compartments
- ✓ 4 scenarios of development:
  - NATural
  - Selective CUtting
  - Legal RUssian
  - ILLegal
- ✓ Simulation results for 200 years (41 time moments)
- ✓ 6 species
- ✓13 age groups

104*4*41*6*13=1,000,000 combinations
For these combinations: 20 attributes!

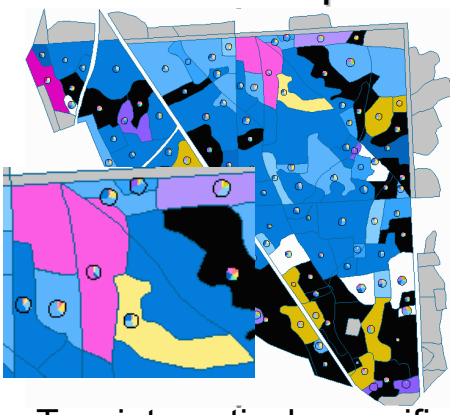
SILVICS - Silvicultural Systems for Sustainable Forest Resources Management
Univ. Wageningen (NL), EFI (FI), AIS (DE), RAS (RU)
INTAS, 2002-2005

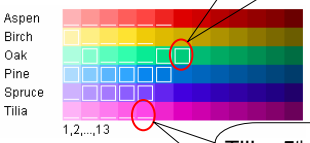
Compare biomass in two scenarios



SCU: rather stable number of forest compartments in all classes; LRU: high temporal variability

Look at biodiversity:
Dominant Species/Age classification



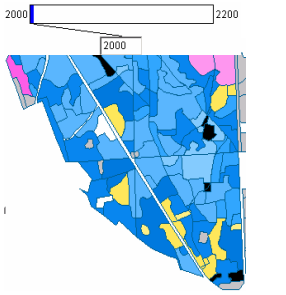


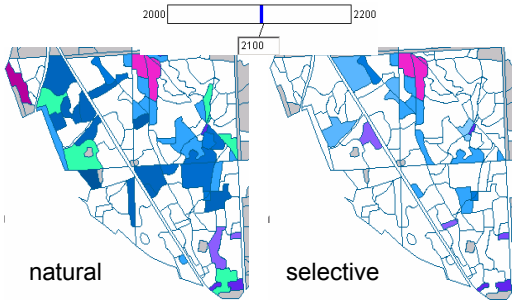
Oak, 7th age group dominates in some compartments

Tilia, 5th age group is present in some compartments

Two interactively specified thresholds
1. Presence
2. Dominance level

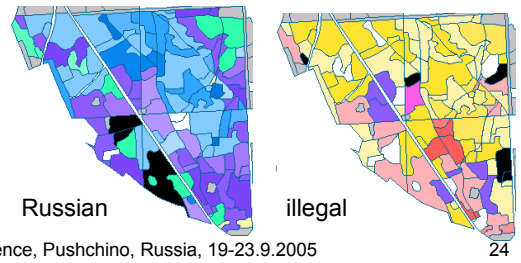
Dominant Species and Age Class (1)





natural

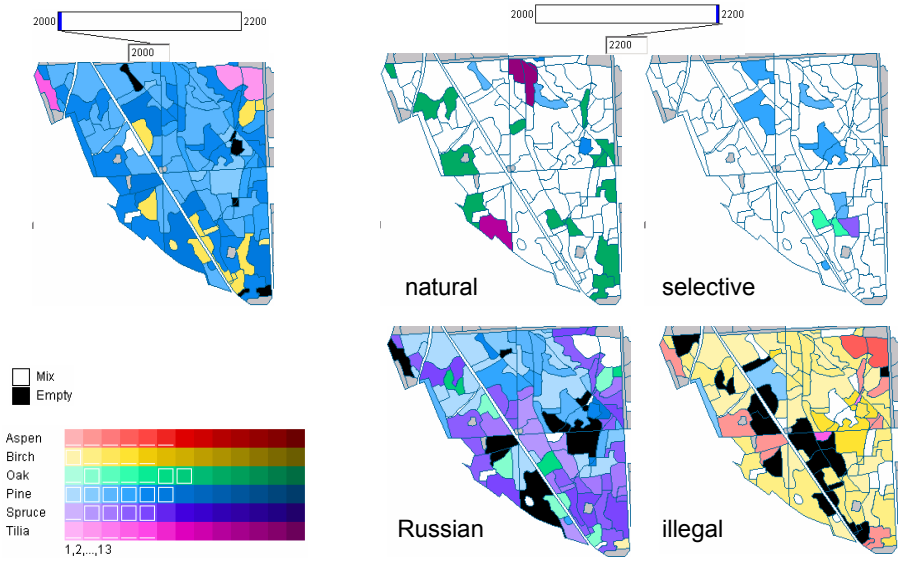
selective



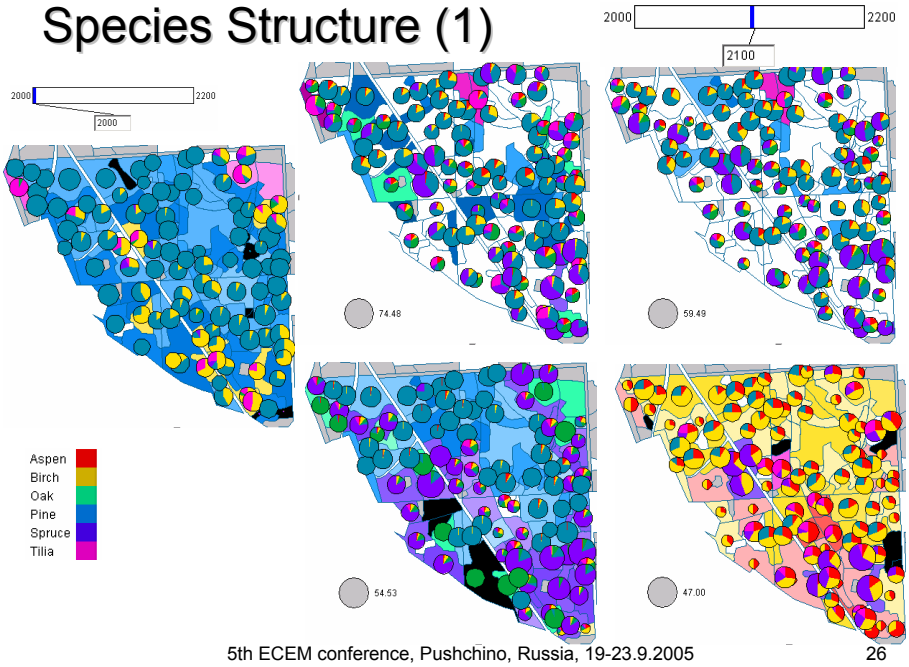
Russian

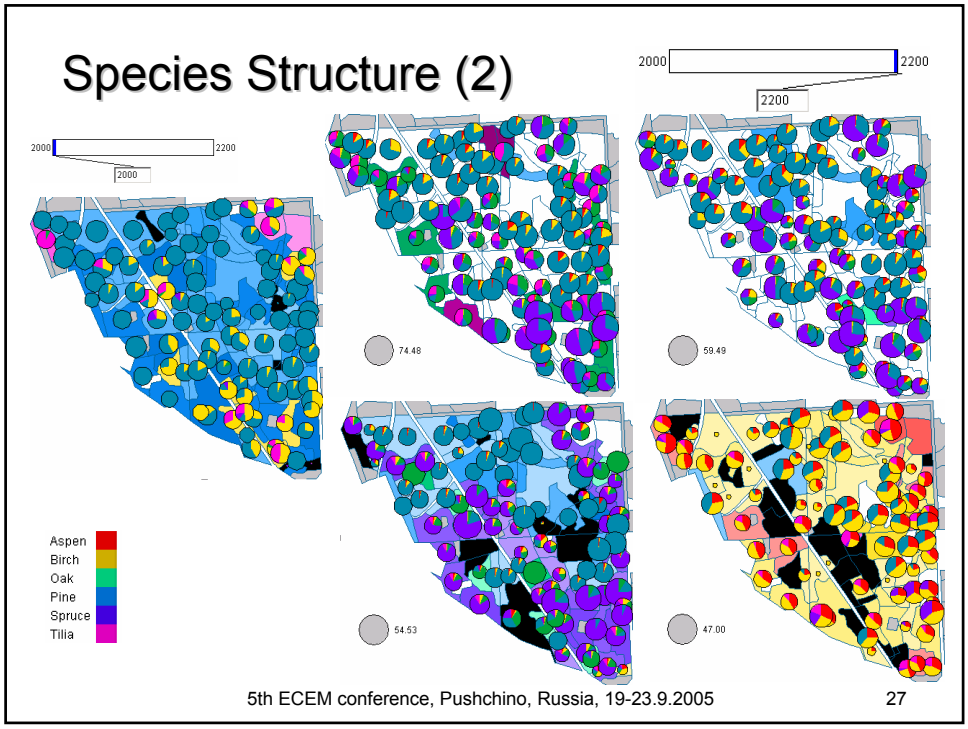
illegal

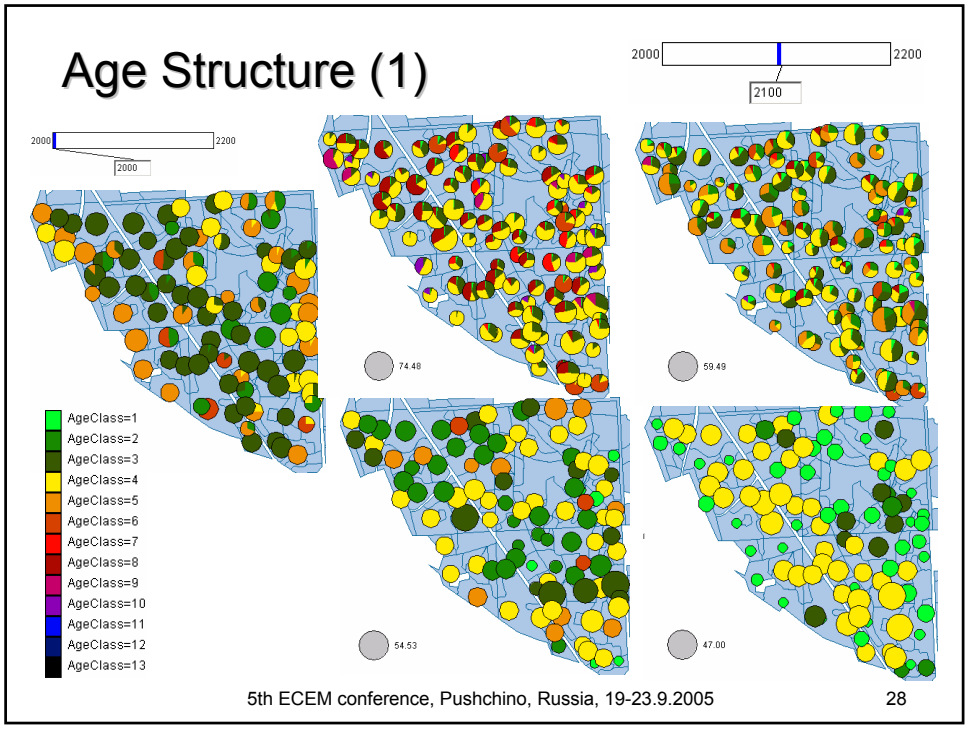
Dominant Species and Age Class (2)



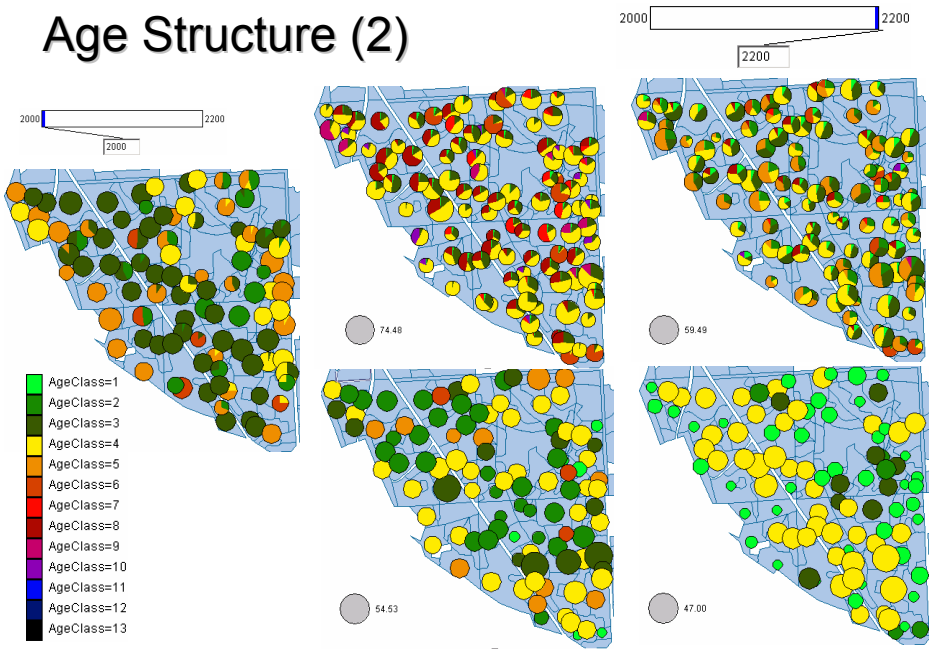
Species Structure (1)




Species Structure (2)
2000
2200
2000
2200
74.48
59.49
54.53
47.00
Aspen
Birch
Oak
Pine
Spruce
Tilia
5th ECEM conference, Pushchino, Russia, 19-23.9.2005
27


Age Structure (1)
2000
2200
2100
2000
2200
74.48
59.49
54.53
47.00
AgeClass=1
AgeClass=2
AgeClass=3
AgeClass=4
AgeClass=5
AgeClass=6
AgeClass=7
AgeClass=8
AgeClass=9
AgeClass=10
AgeClass=11
AgeClass=12
AgeClass=13
5th ECEM conference, Pushchino, Russia, 19-23.9.2005
28

Age Structure (2)



Outline

1. Geo-visualisation's view on ecological modelling: demanding problems and challenging tasks
2. Case study 1: pesticide accumulation
3. Case study 2: forest dynamics
4. A systematic approach to exploratory data analysis (EDA): elements of the general theory
5. Software issues

Recap: aggregation tools

1. Several variants of time series aggregation
2. Aggregation of multiple attributes via selection of the dominant attribute

both in a spatial context

closely integrated with interactive visualisation and data transformation

Roles of aggregation tools in EDA

- Aggregation supports grasping the overall characteristics on the processes / scenarios
- To be instrumental, aggregation tools should be interactive and dynamic for:
  1. Flexible and powerful data transformation
  2. Immediate feedback on visual displays
  3. Analysis of sensitivity to the aggregation parameters
  4. Selection of interesting data instances, access to them
- Intelligent aggregation is important for decision support as a tool for the exploration and evaluation of alternatives

What Is EDA?

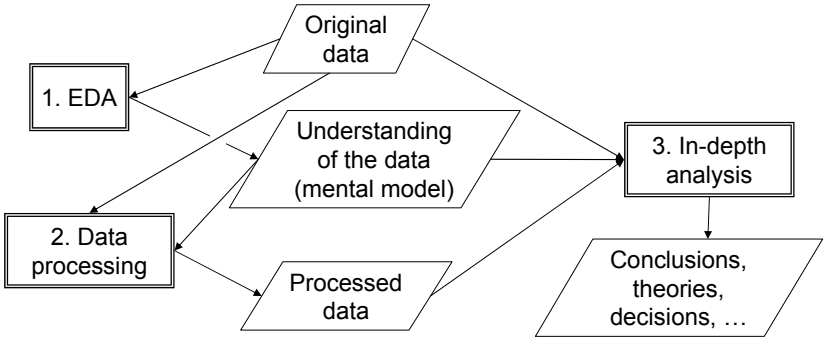
- Emerged in statistics in 1970ies; originator: John Tukey
- A philosophy and discipline of unbiased looking at data: “What can data tell me?” rather than “Do they agree with my expectations?”
  - Similar to the work of a detective (J.Tukey)
- Need to look at data $\Rightarrow$ focus on visualisation and user interaction with data displays

Purposes of EDA

- Uncover peculiarities of the data and, on this basis, understand how the data should be further processed (e.g. filtered, transformed, split into parts, fused, …)
- Generate hypotheses for further testing (e.g. using statistical methods)
- Choose proper methods for in-depth analysis (possibly, domain-specific)

EDA vs. other analyses

- EDA does not substitute rigor methods of numerical analysis, either general or domain-specific, but should give the understanding what methods and how to apply



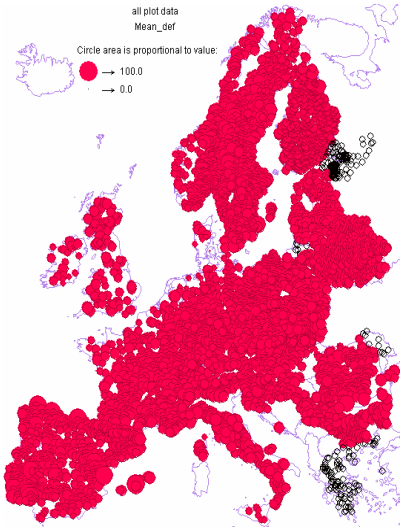
EDA vs. information presentation

- EDA makes intensive use of graphics
- However, “nice” presentation and reporting are not EDA purposes
- Primary goal of presentation: convey certain idea or set of ideas to others
  - Understandably
  - Convincingly
  - Aesthetically attractively
- This requires different visual means than exploration

Case study 3: EDA for the exploration of forest defoliations

- Large volume: 6169 spatially-referenced time series
- Two dimensions: S&T
- Many missing values
- No full compatibility across countries, species, time etc.

Data from NEFIS project



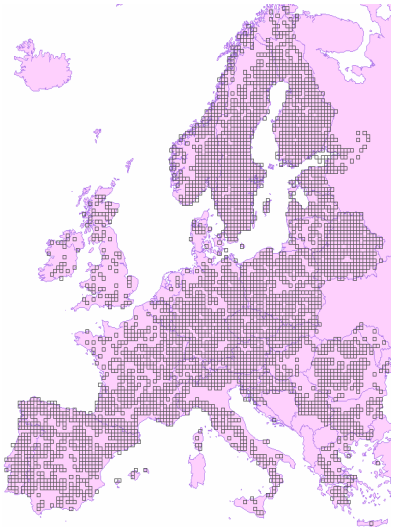
General procedure of the EDA

1. See the whole
   - Space + Time → 2 complementary views
     1) Evolution of spatial patterns in time
     2) Distribution of temporal behaviours in space
2. Divide and focus
   - Data are complex → Have to be explored by slices and subsets (species, age groups, countries, years, …)
3. Attend to particulars
   - Detect outliers, strange behaviours, unexpected patterns, …

See the whole: Handle large data volumes

- General approach: Data aggregation
- Task 1: Explore evolution of spatial patterns
- Appropriate data transformation: aggregate by small space compartments (regular grid with 4025 cells); separately for different species; various aggregates (mean, max)

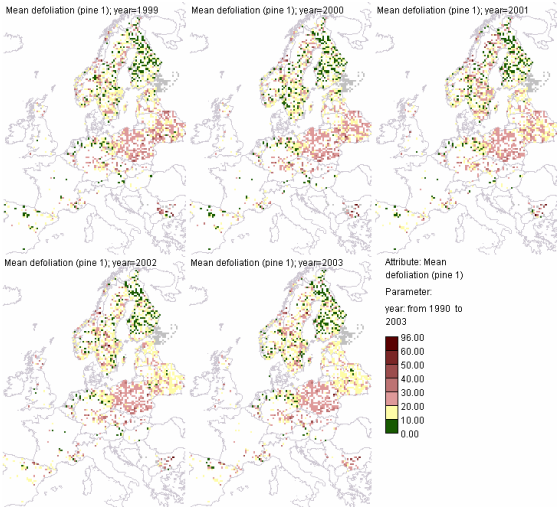
➢ Gain: no symbol overlapping



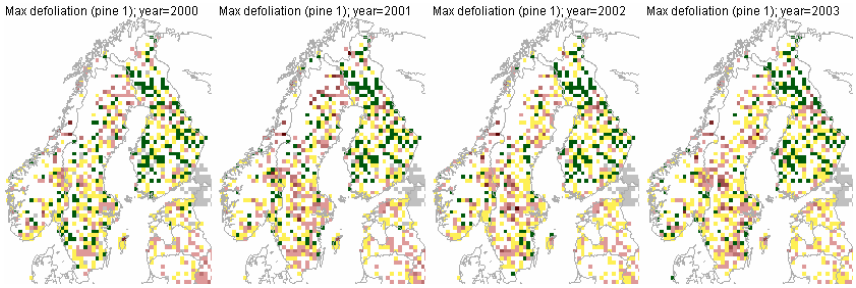
Explore evolution of spatial patterns

a) Animated map
b) Map sequence

Observations:
- Persistently high values in Poland
- Improvement in Belarus
- Mosaic distribution in most countries: great differences between close locations
- Outliers



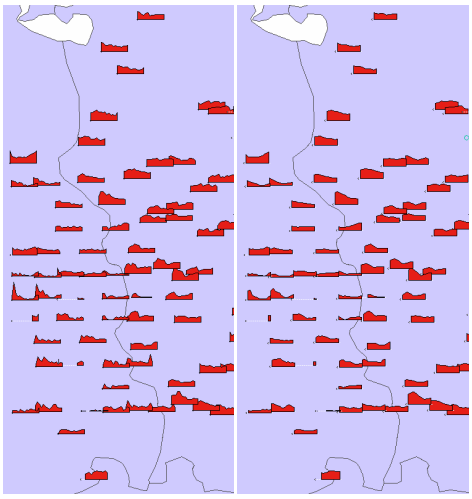
Divide and Focus: Exploration on country level



- Recommendable due to inconsistencies between countries
- Observation: abrupt changes between locations → spatial smoothing methods are not appropriate

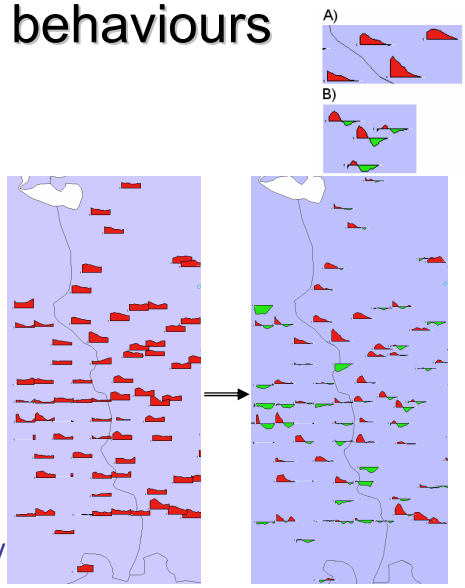
Explore spatial distribution of temporal behaviours

- Are behaviours in neighbouring places similar?
- Step 1. Smoothing supports revealing general patterns and disregarding fluctuations and outliers (we shall look at outliers later)



Explore spatial distribution of temporal behaviours

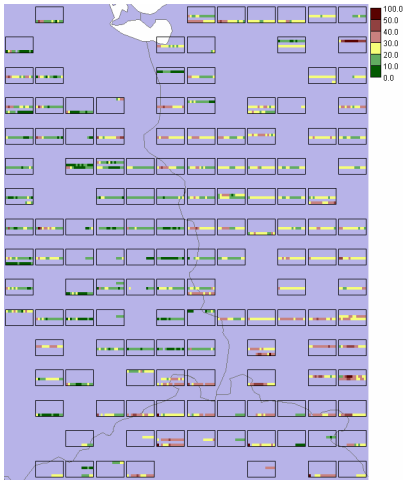
- Are behaviours in neighbouring places similar?
- Step 2. Temporal comparison (e.g. with particular year, mean for a period) helps to disregard absolute differences in values and thus focus on behaviours



Observation: no strong similarity between neighbouring places

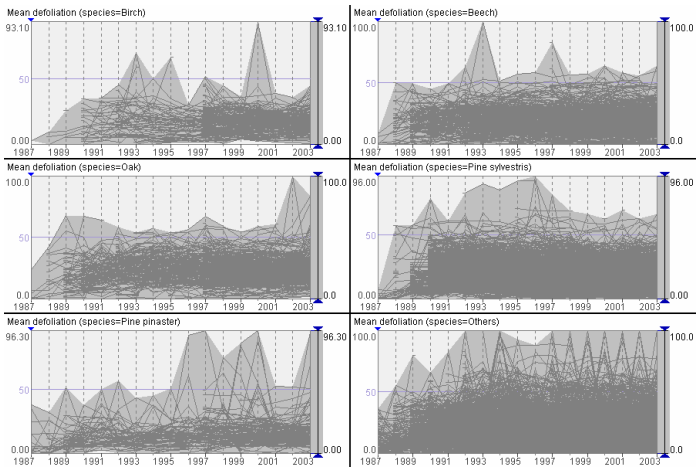
Compare behaviours in plots with different main species

- Mosaic signs:
  - 6 rows for species;
  - 14 columns for years 1990-2003;
  - Colours encode defoliation values



Observation:
behaviours differ for different main species

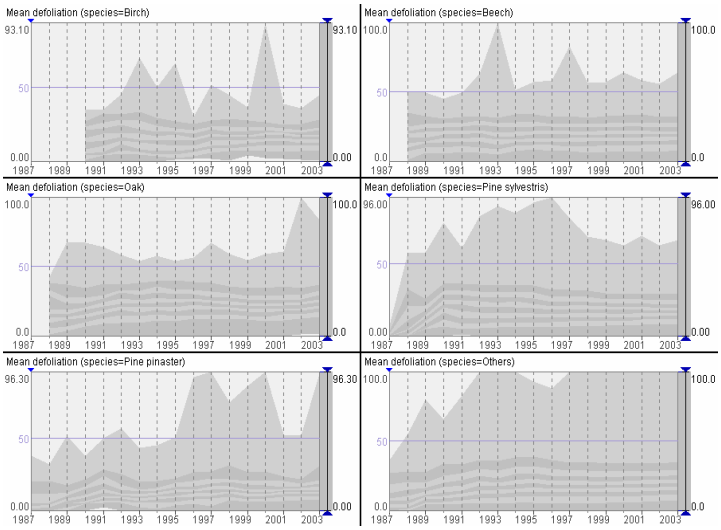
Explore overall temporal trends



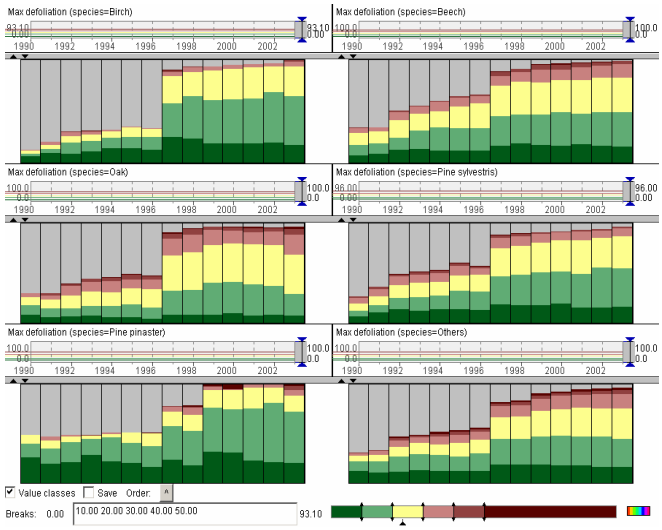
Line overlapping obstructs data analysis

→ apply aggregation

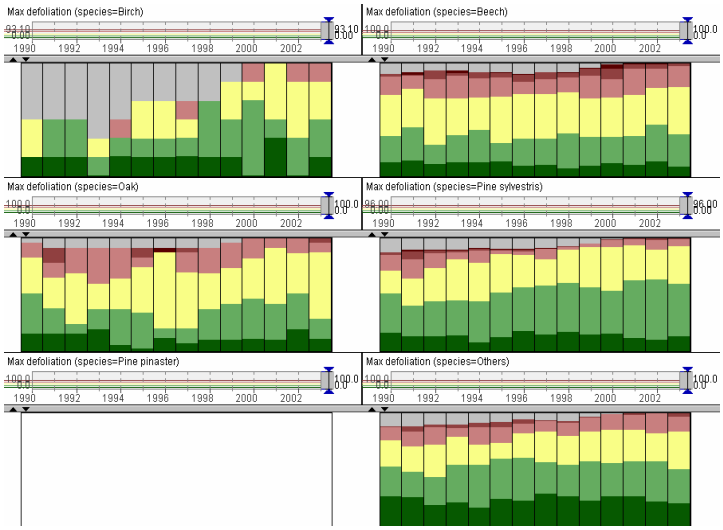
Aggregation method 1: by quantiles



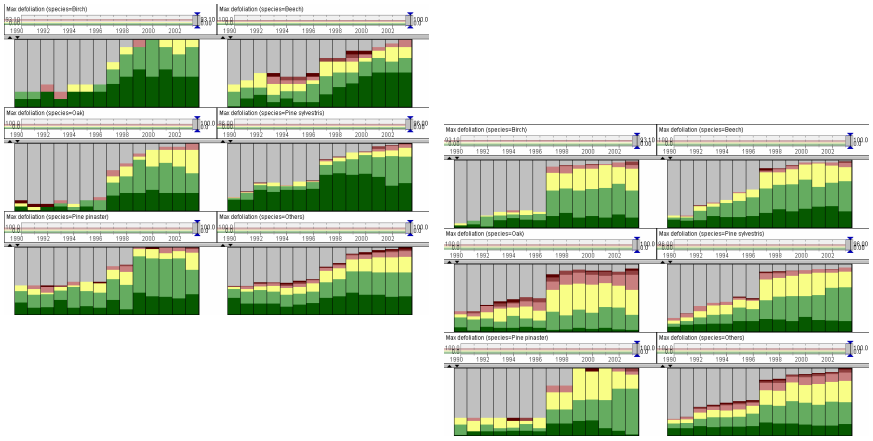
Aggregation method 2: by intervals



Divide and Focus: Germany



Divide and Focus: age groups 1,3



Attend to particulars

Types of particulars (examples):

- Extreme values
- Extreme changes
- High variability
- …

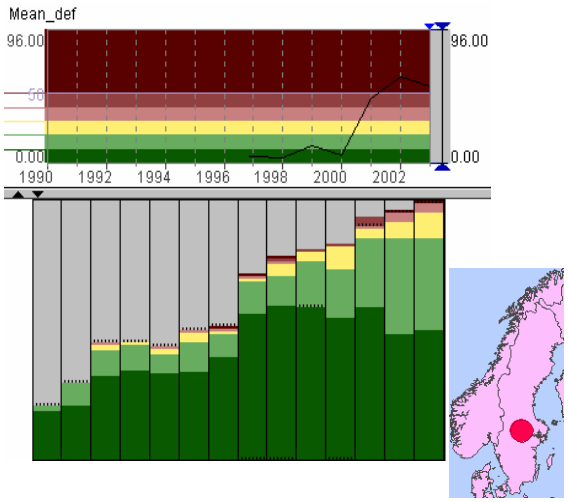
Questions:

- When?
- Where?
- What is around?
- *Why?* (a question for further, in-depth analysis)

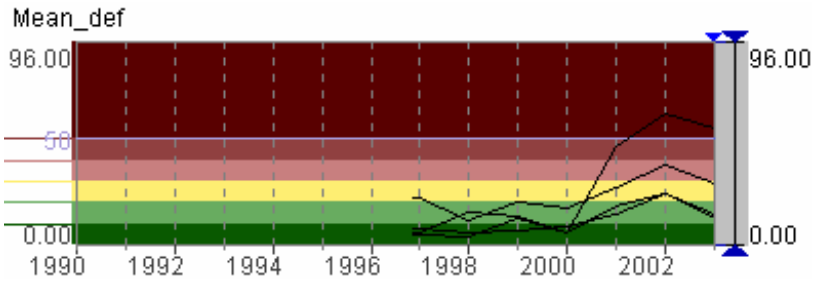
  Domain knowledge is essential

Attend to particulars: extreme values

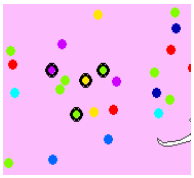
1. Click on a segment corresponding to extreme values
2. The behaviour(s) is(are) highlighted on the time graph
3. The location(s) is(are) highlighted on the map



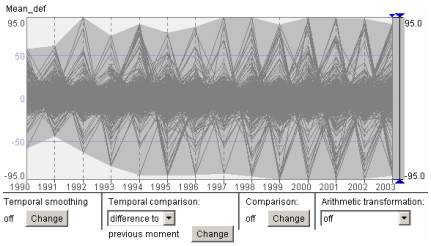
Attend to particulars: what is around?

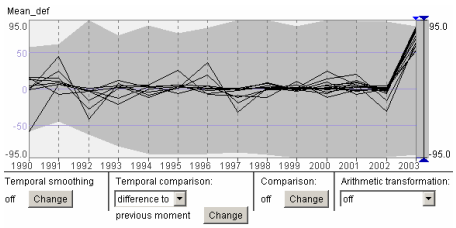


- In some neighbouring places the behaviours during the period 2000 - 2003 are somewhat similar

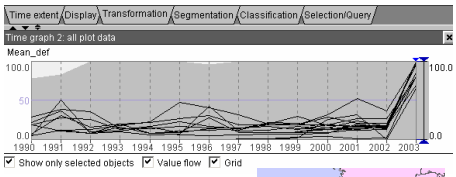


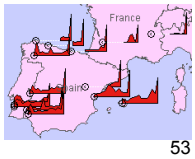
Attend to particulars: extreme changes



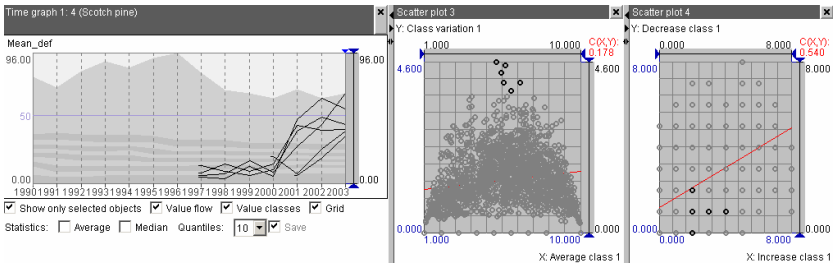


1. Transform the time graph to show changes
2. Select extreme changes in a specific year (here 2003)

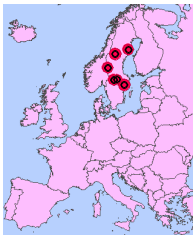




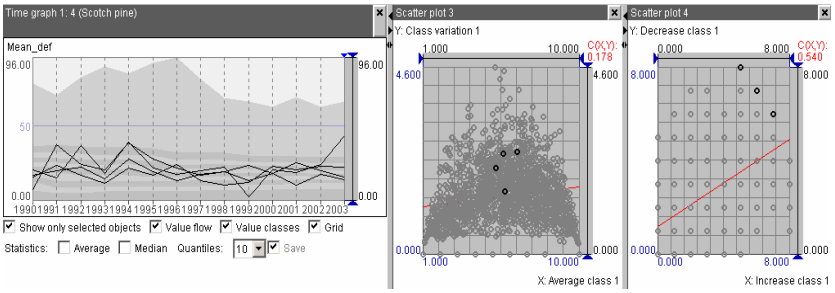
Attend to particulars: high variation



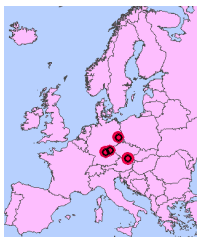
1. Aggregate time graph by quantiles
2. Save counts
3. Visualise e.g. on a scatter plot
4. Select items with high variation



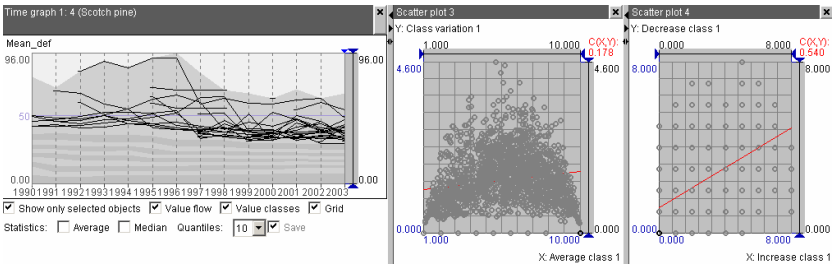
Attend to particulars: high fluctuation



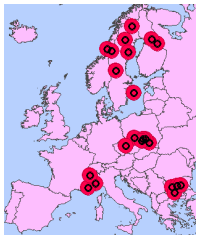
- Select items with maximal number of jumps between quantiles



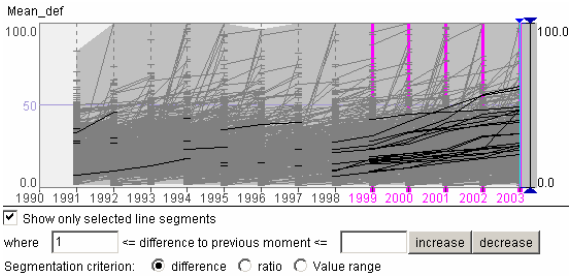
Attend to particulars: stable extremes



- Select items being always in the topmost 10%



Attend to particulars: stable increase



1. Turn the time graph in the segmentation mode
2. Choose “increase” and set minimum difference
3. Select a sequence of years by clicking
4. Check sensitivity to the time period!

Recap: Exploration procedure

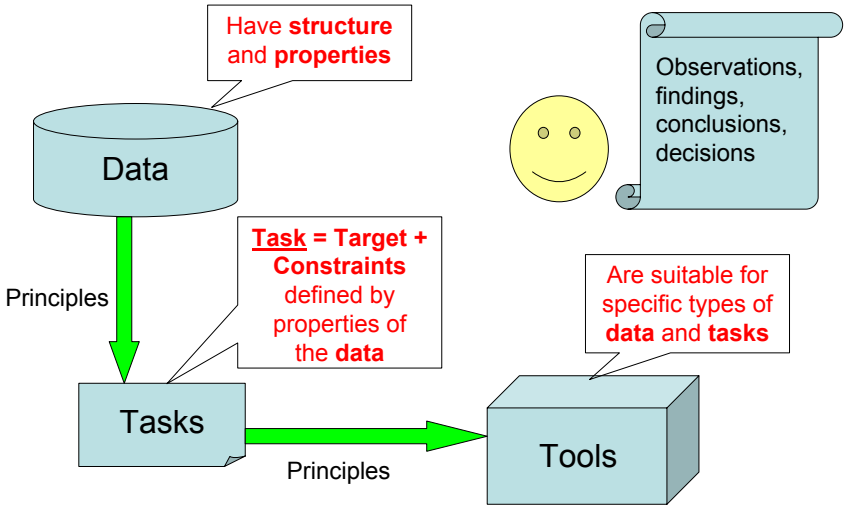
- See the whole
  - Evolution of spatial patterns in time
  - Distribution of temporal behaviours in space
- Divide and focus
  - Data were explored by slices and subsets (species, age groups, countries, years, …)
- Attend to particulars
  - Extreme values, extreme changes, high variation, high fluctuations, stable growth …

Recap: Tools

- Visualisation on thematic maps, time graphs, other aspatial displays
- Aggregation: reduce data volume & symbol overlapping
- Filtering: divide and focus (select subsets)
- Marking: see corresponding data on several displays
- Data transformation: smoothing, computing changes, normalisation etc.

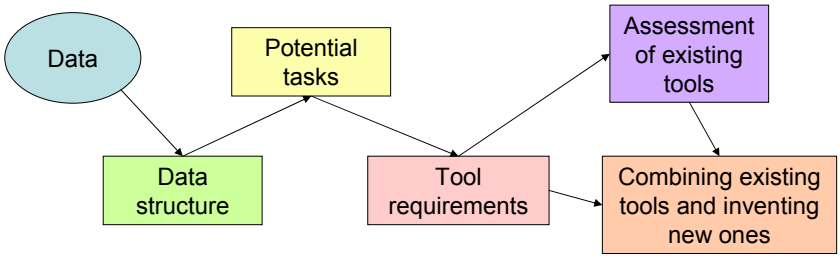
It is important to use the tools in combination

Elements of the theory of EDA



Notes for designers of new tools

✓Tool design (in particular, map design) should base on task analysis!



Outline

1. Geo-visualisation's view on ecological modelling: demanding problems and challenging tasks
2. Case study 1: pesticide accumulation
3. Case study 2: forest dynamics
4. A systematic approach to exploratory data analysis (EDA): elements of the general theory
5. Software issues

Requirements to EDA software

- Space- and Time-awareness
- Work with complex multidimensional data
- Support for uncertain and missing data
- Scalability
- Support and encouraging of several complementary views on the same data
- Dynamic linking and coordination of several data displays
- From the overall view to particulars of interest
- From idea generation to hypothesis testing using statistical methods, followed by reporting

GIS for EDA: major problems

- Time-awareness
- Work with complex and multidimensional data
- Processing uncertain and missing data
- Scalability
- Interactivity of visualisations
- Dynamic linking of multiple displays
- Idea processing

Potentially useful tools for EDA

- Information visualisation tools, for example, HCE & TimeSearcher from HCIL, Univ. Maryland
- Geovisualisation tools, for example GeoVistaStudio (Penn State Univ.) and Descartes/CommonGIS (Fraunhofer Institute AIS)
- Graphical statistics tools, for example, Manet & Mondrian (Augsburg Univ.)
  - ☞ Usually such systems are research prototypes that implement innovative ideas, but provide restricted functionality and limited user support

CommonGIS (not a “common GIS”)

A variety of well-integrated tools for EDA

- Time-aware maps + statistical graphics; several mechanisms of display coordination
- Designed to gain synergy of
  - ✓ Visualisation
  - ✓ Display manipulation
  - ✓ Data manipulation
  - ✓ Querying
  - ✓ Computational techniques, including aggregation and data mining

- *Quick demo?*

Still open issues (for all tools!)

- Work with qualitative (non-numeric) data
- Work with fuzzy, uncertain, and missing data
- Continue scalability efforts
- Intelligent guidance through the overall process of data analysis, avoiding cognitive complexity
- Adaptability to user, data, tasks, and hardware
- Support in processing and management of observations: recording, structuring, browsing, searching, checking, combining, interpreting…
- Help in visual communication of derived data, constructed knowledge, and recommended decisions

Conclusions

1. EDA is essential in ecological modelling for preparation of data, verification and tuning of models, interpretation of results, and evaluation of decision alternatives
2. Systematic application of EDA requires careful consideration of characteristics of data, relevant analytical tasks, properties of tools
3. EDA tools should combine interactive visualisation with data transformation, dynamic query, and sophisticated computations
4. Still there are many things to do… for scientists and for software developers

To Learn More:

International Cartographic Association
Commission on Visualization

- Software: http://www.commongis.com
- Papers, tutorials, on-line demos: http://www.ais.fraunhofer.de/and
- Book to appear:

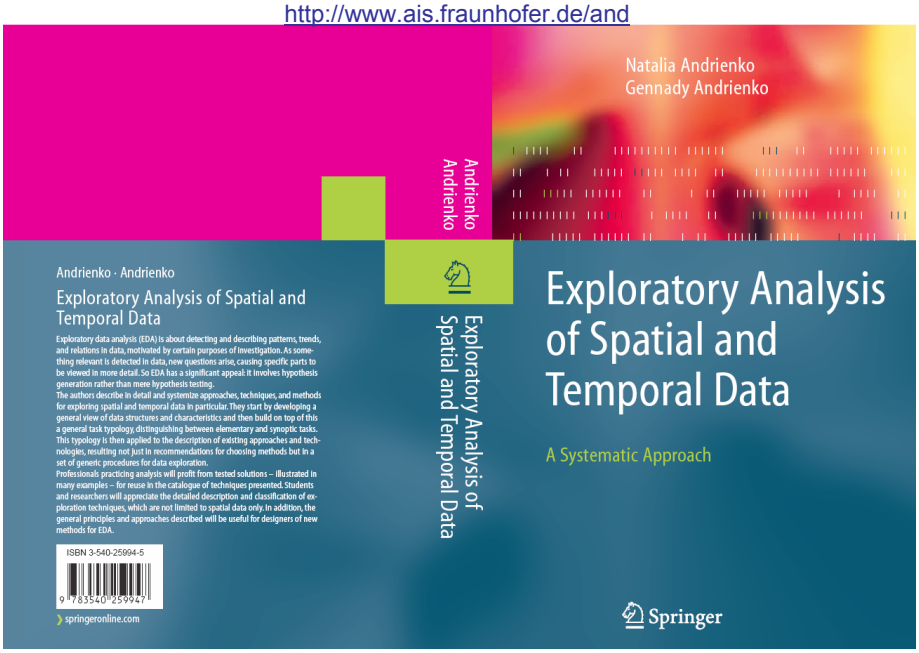
Natalia and Gennady Andrienko

"Exploratory Analysis of Spatial and Temporal data. A Systematic Approach"

(Springer-Verlag, ≈ end 2005)

A theoretical framework for linking tasks, tools, and principles of data analysis

http://www.ais.fraunhofer.de/and



In press, to appear ≈ end 2005