

IT Tools for an Integrated Data Editing Concept

- The story so far ...
- Draft of Methodological Considerations
- Module: Selective Data Editing
- Module: Automatic Imputation
- Scheduled Enhancements in Automatic Imputation

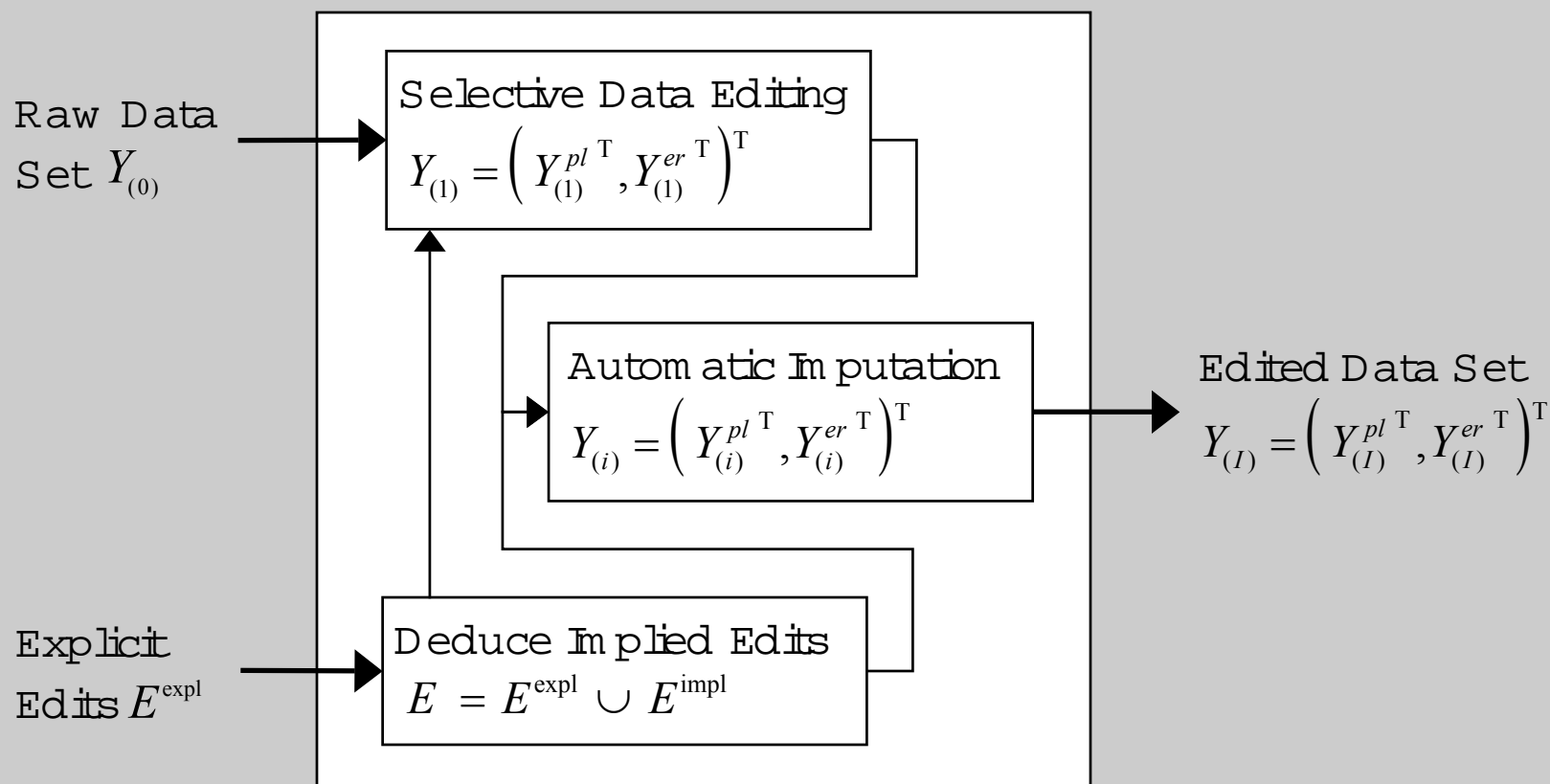
The Story so far ...

- **Initial point:** Formulating the methodological requirements for data editing and modelling the data editing processes within the statistical offices in Germany

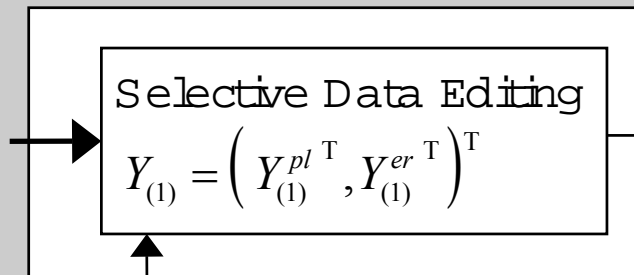
- **Implementation step 1:** Developing IT-tools for planning and initialising data editing processes
 - DE-Intranet Pages as an online reference for the application flow and the applicable methods during data editing
 - DE-Schedulers for rough and detailed planning of data editing
 - DE-Editor for the specification of data set descriptions and editing rules (with access to the specifications of previous and similar surveys)

- **Implementation step 2:** Implementing methods and algorithms for data editing processes

Draft of Methodological Considerations

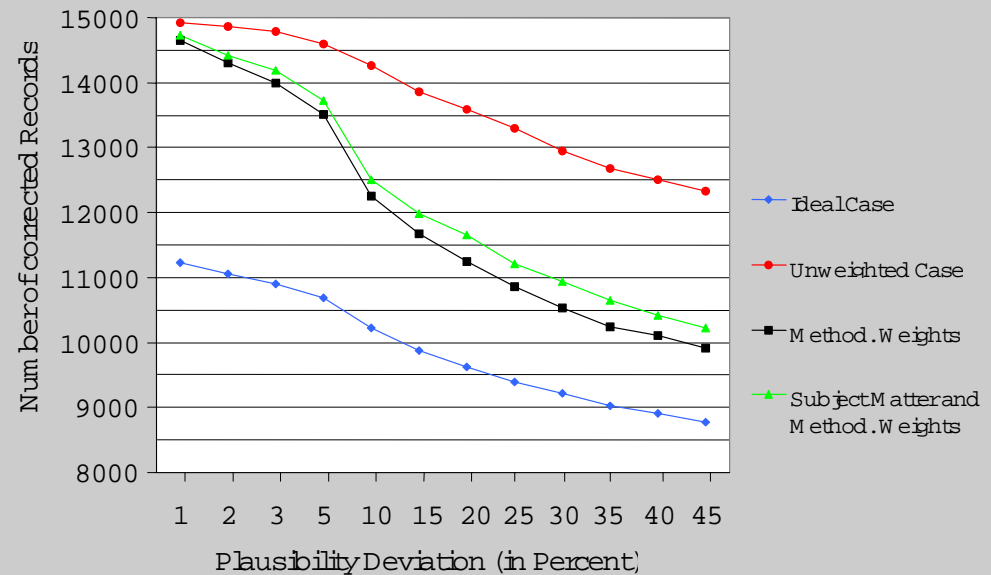


Module: Selective Data Editing



- Implementation: SAS Macro Library
- Ongoing improvements (Nearest-Neighbour Approach, adaptation to other surveys, ...)

- Two simulation studies with annual Cost Structure Surveys
- Example: Efficiency graph in producing industry survey from 2000



Module: Automatic Imputation



```

for (i = 1 ; i <= I && Y_{(i)}^{er} ; i++) {
  (1) Create Y_{(i+1)}^{er} by evaluating and converting the
      (i+1)-th error configuration in Y_{(i)}^{er}.
  (2a) Generate monotone missing pattern by applying
      SAS PROC MI on Y_{(i+1)} = (Y_{(i)}^{pl T}, Y_{(i+1)}^{er T})^T.
  (2b) Impute missing values by applying IVEware on Y_{(i+1)}.
  (3) Create Y_{(i+1)}^{pl} by transferring completed observations
      from Y_{(i+1)}^{er} that pass E to Y_{(i)}^{pl}.
}
  
```

Scheduled Enhancements in Automatic Imputation

- Implement SAS-M across for evaluating the i -th error configuration

- Replace IVEware by tailor-made SAS-M across ...
 - ...providing the original functionality of the basically linear approach
 - ...adding frequentistic model selection with the Approximative Bayesian Bootstrap
 - ...extending the linear approach by Multilayer Perceptrons and Support Vector Machines that provide non linear and more general models
 - ...executing external C-Routines for CPU-intensive work steps

- Carry out extensive simulation studies both on
 - production data from completed surveys (cost structure surveys and micro data)
 - simulated data sets with controlled Missing Mechanism