TALL SOM FORTELLER

Data Quality Dashboard

TRYM LANGBALLE – STATISTICS NORWAY

Statistisk sentralbyrå
Statistics Norway

Agenda

Motivation

Quality indicators

Robust variance estimates

Outlier detection

Demo - Graphical User Interface

Motivation

- Need for feedback on quality of newly established surveys of service industries
- Develop a unified system for assessing data quality
- New Python production system for currently SPPI being developed

Quality Indicators

Recomended Quality Indicators in Official Statistics (SSB 2024/5)

- Comparable to
  - Single Integrated Metadata Structure (ESS)
  - Generic Statistical Business Process Model (UN ECE)
- Chosen indicators for the dashboard:
  - Imputation rate
  - Data completeness
  - Unit response rate
  - Uncertainty (Coefficient of variation)
  - Control failure rate


Statistisk sentralbyrå
Statistics Norway

Anbefalte kvalitetsindikatorer i offisiell statistikk

TALL SOM FORTELLER

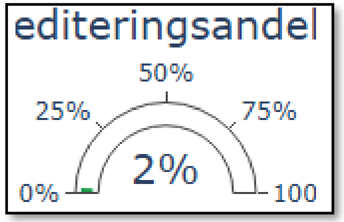
Aslaug Hurlen Foss, Per Ola Haugen

NOTATER / DOCUMENTS

2024/5

Imputation rate

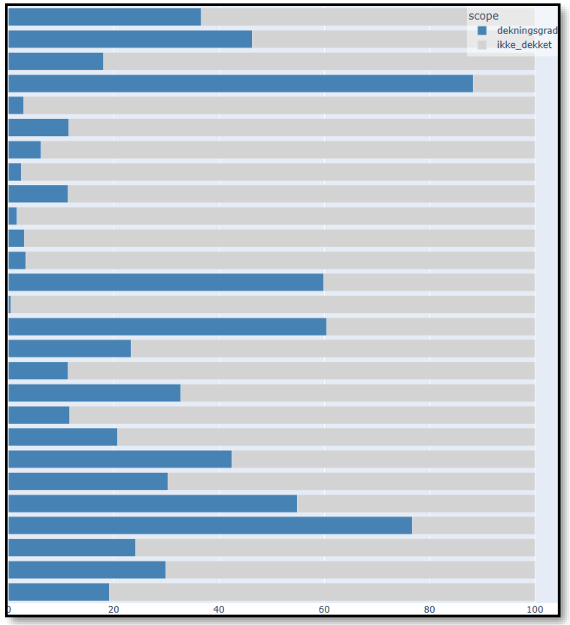
- Share of prices that has been changed. Either through manual correction or an automatic process.
- Requires that all individual data cells (price) must have associated process data about how it has been changed
- Interpretation : A high imputation rate might be an indicator that the accuracy of the recieved data is poor.
- Visual representation:



Statistisk sentralbyrå
Statistics Norway

Data completeness

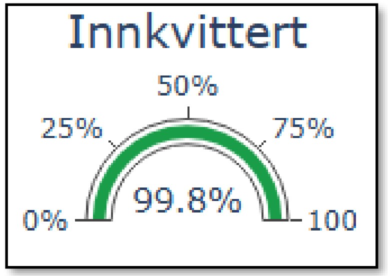
- Share of units with value (non missing) in relation to number of units in the data set.
- Target: Idealy we want the share of units with values to be as close to one as possible.
- Visual representation:



Statistisk sentralbyrå
Statistics Norway

Unit response /Item response

- Share if units recieved in relation to the total units in data sett
- Target: Idealy we want the share to be as close to one as possible.
- Visual representation:



Statistisk sentralbyrå
Statistics Norway

TALL SOM FORTELLER

Uncertainty
(Coefficient of variation)

Models for elementary indices

- Each Index can be expressed with a regression model
- Prices are random observations of an underlying inflation rate
- Carli: $I_{ij}^{s,t} = \frac{p_{ij}^{t}}{p_{ij}^{s}} = \theta_i + \epsilon_{ij}$
- Dutot: $p_{ij}^{t} = \theta_i \, p_{ij}^{s} + \epsilon_{ij}$
- Jevons: $\log I_{ij}^{s,t} = \mu_i + \epsilon_{ij}$

Statistisk sentralbyrå
Statistics Norway

Variance

- Theoretical variance derived under assumptions like homoscedasticity (equal variances for all price relatives).
- Variance tells us how much individual prices disagree:
  - If all move together → variance = 0.
  - If prices move very differently → variance is larger.
- Example (Carli):

$$V(P) = \frac{\sigma_i^2}{n}$$

Plug-in Estimation

- Idea: Estimate variance from residuals and substitute into formula.
- Example (Carli):

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum \left(I_{ij} - \widehat{\theta}_i \right)^2$$

$$V(P) = \frac{\hat{\sigma}^2}{n}$$

- Works if all products share the same variance

Problem: Pricerelatives does not have same variance

- Theoretical formulas assume prices have the same variance.
- In reality, some products are more volatile than others.
- Robust method:
  - Look at the actual residuals (differences between observed and average).
  - Use them to estimate variance without strong assumptions.
- More realistic for messy real-world data?

Robust variance estimates (Zhang, 2010)

- Carli: $\hat{V}\left(P_i^{s,t}\right) = \frac{\sum_j e_{ij}^2}{n_i(n_i - 1)}$

- Dutot: $\hat{V}\left(P_i^{s,t}\right) = \left(\sum_k p_{ik}^s\right)^{-2} \sum_j \left(1 - \frac{p_{ij}^s}{\sum_k p_{ik}^s}\right)^{-1} e_{ij}^2$

- Jevons: $\hat{V}\left(P_i^{s,t}\right) = \left(P_i^{s,t}\right)^2 \frac{\sum_j e_{ij}^2}{n_i(n_i - 1)}$

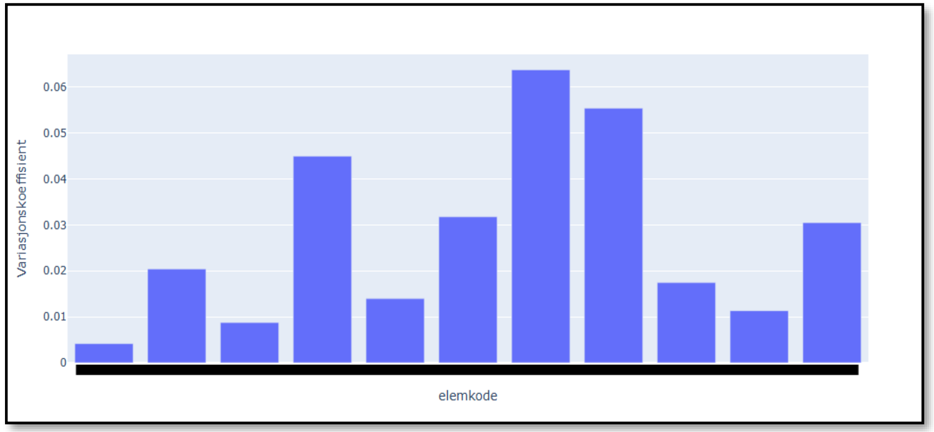
- Variance estimates are now only dependent on the residual

Coefficient of variation

- Relative measure of uncertainty
- Easier to compare variances between different service areas

$$CV = \frac{\sqrt{var(\hat{\theta})}}{\hat{\theta}}$$

- Visual representation:



Control failure rate

- Measurement: Share of units failing checks in relation to the number of units checked
- In this context, we consider outliers outside a threshold value to be a failure.
- Two ways used calculate outliers:
  - Studentized Residuals - R_Student
  - Difference in Fit Studentized – DIFFITS

Statistisk sentralbyrå
Statistics Norway

Control failure rate continued

- Calculated using a simple linear regression model:

$$P_{i,t} = \beta P_{i,t-1} + \epsilon_i$$

$$\widehat{\epsilon_i} = P_{i,t} - \beta P_{i,t-1}$$

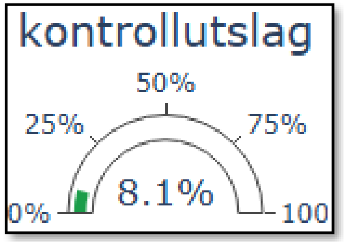
- Studentized Residuals:

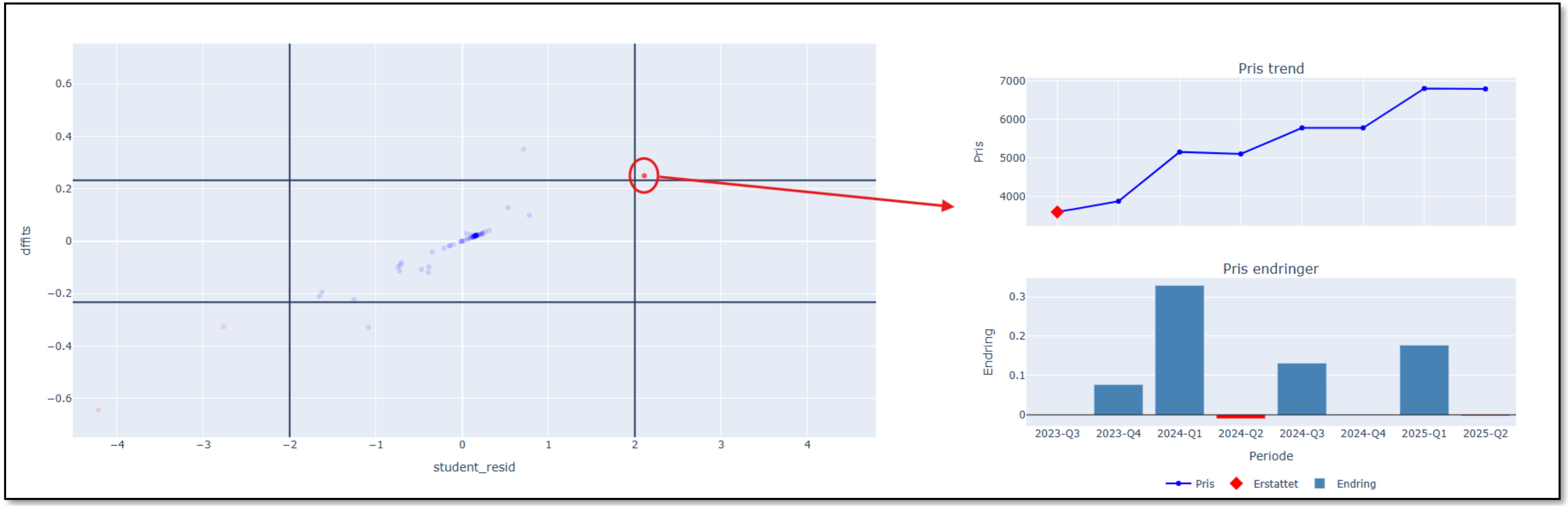
$$\frac{\widehat{\epsilon_i}}{\sigma\widehat{\sqrt{i - h_{ii}}}} > 2$$

- Studentized Difference in fit:

$$\frac{\hat{y}_i - \hat{y}_{i(i)}}{s_i\sqrt{h_{ii}}} > 2\sqrt{\frac{1}{n}}$$

Visual representation





Statistisk sentralbyrå
Statistics Norway

Graphical User Interface

- Interactive dashboard for monitoring the quality indicators of our sample
- Background written using Python
  - Main programming language at Statistics Norway
  - Open source
  - Some SQL queries
- Dashboard made using the Dash framework
  - A package for making data-dashboards in Python
  - Some HTML and CSS

Data sources

- SQL data tables containing metadata on units in sample
- Price data mainly from surveys (other data sources are also represented)
- Production file – containing information on: type of imputation, base prices, codes for aggregation etc.

TALL SOM FORTELLER

DEMO!

Statistisk sentralbyrå
Statistics Norway

TALL SOM FORTELLER

Thank you!

Statistisk sentralbyrå
Statistics Norway