UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Expert Meeting on Statistical Data Editing

3-7 October 2022, (virtual)

Experimental Short-Term Statistics based on Data Imputation Methods

Jan Ditscheid (Federal Statistical Office, Germany)

jan.ditscheid@destatis.de

I. Introduction

1. The recent COVID-19 pandemic has catapulted reliable data on current economic developments to center stage of academic and public debate. Therefore, the Federal Statistical Office's "t+15" project aims to improve statistical production processes to provide additional critical economic indicators much faster than before, at t+15 (currently, the indicators are published between t+30 and t+70). Since data availability is limited due to reporting delays at the time targeted, experimental data or nowcasts are used as proxies for the final data releases. This work reviews statistical imputation methods as one possible approach to deal with missing values to accelerate data availability. Examples include univariate techniques, donor-based procedures, and regression-based machine learning methods. Based on historical data, the suitability of the methods for the subsequent estimation is assessed. In a final step, those procedures with significantly good results are used to estimate real-time data.

2. In chapter 2, we provide an overview of the data source and transformation patterns. In chapter 3, we dwell on the main indices to be estimated within the project. In chapter 4, we deal with various imputation methods and provide a comparative analysis of the best performing one, with regard to the performance measures outlined in the same chapter. We aim to further optimize the best performer to improve estimates of our short-termed statistics in chapter 5. Chapter 6 concludes.

II. Data Source

A. Origin of the Data

3. The data used to calculate the turnover index and the new orders index are based on the monthly report for companies in the manufacturing sector. The statistical offices of the federal states collect monthly data on a decentralized basis.

B. Structure of the data

4. Firm-Level based data is available from January 2018 until today. Variables available include firm number, firm type, time of data entry, reference year, reference month, state, industry at 4-digit, 3-digit and 2-digit levels, employee size classes, number of employees, domestic sales, foreign sales, foreign sales (non-EU), domestic incoming orders, foreign incoming orders and foreign incoming orders (non-EU). Reports are either complete or not available. Therefore, in case a firm has not reported until t-15, only firm-specific values from the past are available to impute variables of the current reporting period. Before we move to real-time applications, our procedures are tested using historical data.

C. Preparation of the Data

5. For historical data, all reports have been available since 2018. However, for each reporting month to be estimated, reports that were not present at t+15 are removed, and thus the situation as it exists at the current margin is simulated. For the removed reports, new reports are created that contain only that information available at "t+15". For these reports, the missing values are imputed.

III. Indices to be estimated

A. Turnover Index

6. Turnover comprises the value of all own products and industrial/craft services supplied to third parties in the reference month by firms with 50 or more employees in the manufacturing sector. The survey distinguishes between sales from the domestic market and those made to foreign recipients.
The Turnover index is a weighted average of domestic, foreign, and non-EU sales for each industry 4-digit. The Federal Statistical Office provides preliminary results approximately five weeks after the reporting month's end, that is, t+35.

B. New orders index

7. New orders comprise the value (excluding sales tax) of all orders for the delivery of self-manufactured (or subcontracted) products firmly accepted in the respective reporting month by firms with 50 or more employees in the manufacturing sector. For domestic and export demand analysis, a distinction is made between orders received from domestic customers and orders placed by foreign customers. The new orders index is a weighted average of domestic new orders, foreign new orders, and foreign new orders (non-EU) for each industry 4-digit. The Federal Statistical Office provides preliminary results approximately five weeks after the reporting month's end, that is, t+35.

IV. Pre-experiments

A. Quality criteria used

8. To quantifiy the quality of our imputations methods the following quality criteria are used.
The Mean Absolute Error (MAE) is defined as (Hyndman and Athanasopoulos, 2021):

$$MAE = \frac{1}{T}\sum_{t=1}^{T} |y_t - \hat{y}_t\,|$$

and calculates for a sample of n (here: the number of periods considered) the average value of the absolute difference between the true value (in our case: the published index y_t from period $t,$ for $t = 1, \ldots$,T, which was built also on data reported later than t+15) and its model prediction. Model prediction in our case means that the index is predicted ($\hat{y}_t$) by using reported data until t+15 and imputing the data of those firms that did not report until t+15 within an interesting period t.
The MAE indicates how far the estimate and the actual value differ on average. The lower the MAE, the better the estimate. This criterion is a reference to the quality measurement of official statistics. Eurostat uses the mean absolute revision as a revision measure (Eurostat, 2014).

9. The RMSE (root mean square error) is calculated as follows:

$$RMSE = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(y_t - \hat{y}_t)^2}$$

The root mean squared error also measures how far the estimated values deviate from the actual values. Squaring gives greater weight to larger deviations (Hyndman and Athanasopoulos, 2021). The lower the RMSE, the better. The RMSE is also used by Eurostat, among others (Eurostat, 2014).

10. When investigating short-term statistics, it is essential to ensure that the direction of economic development is adequately reflected. In order to incorporate this quality measure of the reliability of an estimate, the percentage of failed estimated signs is used as a third quality criterion:

$$POWS = a/T * 100$$

where *a* is the number of failed estimated signs and *T* is the number of estimated periods. POWS is the abbreviation for "Proportion of wrong signs". The lower the POWSC, the better.

B. Tested methods

11. A distinction is made between univariate, tree-based, donor-based, and regression-based methods. Univariate methods only use the information provided by the t variable of investigation itself to estimate the imputed values. Tree-based methods make decisions at the various decision nodes. Donor-based methods use an observed value of the target variable of another observation unit for a missing value. Regression-based methods use regression models to estimate the target variable.

C. Results

12. In our preliminary experiments, the regression-based imputation performed best in terms of the mean absolute error and the root mean square error. Based on these results, this process is optimized in our finetuning process. All tested procedures are applied with the R-Packet “mice”. There are several regression models in the R package. The function "linear regression, predicted values" is applied. This means a linear regression using a least squares’ estimation is used and the values predicted by the model are imputed (Abdalla et al., 2022).

D. Preliminary considerations for optimization

13. To specify the optimal model the correct explanatory variables must be selected. No variables from the current reporting period are available for the unreported observations. In this case, variables that do not change over time can be considered. This allows us to infer some variables from the current reporting period. Industry at the 4-digit level is an essential variable in the calculation of the sales index and the new orders index because the variables of interest are aggregated at the 4-digit level and then a weighted average is calculated. Furthermore, we check which variables from a firm’s previous period can explain the firm’s target variable from the current period most precisely. In our case, these are the federal state, the economic sectors and the unit type. In order to achieve the best possible results at the industry level, it is useful to learn different models for different industries. Here, we assume that different linear relationships exist in various economic sectors. There is the possibility to do so on a 2-digit, 3-digit or 4-digit level. In addition to utilizing the past data of the corresponding firm, it is also possible to use information obtained from the observations already reported from the reporting period until t+15. For example, growth rates can be used to reflect the current economic situation. It is assumed that the growth rates for the observations reported at t+15 roughly corresponds to the growth rates for all observations that have to report this month. This assumption is necessary to justify why we also use the rate of change that we calculate only on the basis of the already existing values to estimate the still missing observations. In addition, it may be helpful to add a variable that maps time-fixed effects for each month. It is also examined whether there are systematic differences between companies that report early and companies that do so later. Thus, no matter whether a company has reported or not, we can utilize information for all companies in the current reporting month.

14. Linear regression models have potential singularity issues due to multicollinearity. Here it is necessary to investigate the various combinations of explanatory variables.

V. Optimization of the models

15. First, we look at all the potential explanatory variables. The state, unit type, and industry 4-digit are nominally scaled, and therefore no correlation (nor rank correlation) can be calculated for these variables.
Usually, the correlation between numerical variables is calculated. Between at least ordinally scaled variables a rank correlation can be calculated after all. Since predominantly categorical variables are examined, a classical correlation analysis is omitted here. For this reason, we build models consisting of only one target variable and one explanatory variable. We then investigate the coefficients of determination to see which variables can explain the target variable significantly.

A. Relationships between explanatory variables and the variables needed for the new orders index

16. To identify which variables can best explain the target variable, a model is learned in which the target variable is explained with only one regressor. After these models have been learned, the coefficients of determination of the models are compared. The rows contain the regressors and the columns contain the target variables.

	Domestic incoming orders	Foreign incoming orders	Foreign incoming orders (non-EU)
Type of Unit	0,0106	0,0083	0,0057
Federal state	0,0026	0,0017	0,0013
Industry 4-digit	0,0835	0,0796	0,0690
Turnover (domestic/foreign/non-EU) of the previous month	0,6192	0,3785	0,6644
Incoming orders (domestic/foreign/non-EU) of the previous month	0,6095	0,8021	0,7727
employees of the previous month	0,4614	0,5388	0,4613

It is unclear which variables can be combined to explain the target variable most precisely. It can be seen that state, industry 4-digit and type of unit solely do not explain new orders well. However, these variables are constant over time and help to classify firms. Reasons can be found both for including the variables and for omitting them. For causality reasons, it is only analysed how well the domestic new orders of the previous month explained the domestic new orders of the current month and how well the foreign new orders of the previous month explained the foreign new orders of the current month. For the time being, domestic orders are not compared with foreign orders. With the help of the number of employees of the previous month, the order intake can be estimated much better than with the variables mentioned before. The number of employees is also relatively constant over time. The constancy over time is lower than for the previously mentioned variables but also relatively high. Relying exclusively on strongly fluctuating variables from the previous month includes the risk that the adjustment to the last month will be too high. Therefore, it may be helpful to include at least a few constant variables. The previous month's new orders can reasonably well estimate the current month's new orders. The correlation between order intake in the current month and the last month could be more significant if order intake were not subject to repeated major fluctuations due to large orders. The industry 4-digits "Shipbuilding" and "Boat and yacht building" are extreme examples of this. In July 2021, aggregate domestic new orders in these industries were over 3.1 billion and fell below 45 million in August 2021. Such fluctuations in some sectors can only be explained by large orders. It is still unclear whether a different approach should be taken for industries with such fluctuations than for industries where new orders are fairly constant over time. The new orders and the previous month's turnover are the variables best suited to explain the new orders with only one variable. It is crucial to find the optimal combination of variables. As mentioned in the previous reflections, it is also helpful to check which variables not present in the original data can be created to improve the models further.

17. In the following, it is analyzed if enough reports to learn separate models for each industry 4-digit are available.

Industry 2-digit	Number of observations
Coal Mining	501
Extraction of crude oil and natural gas	805
Ore mining	51
Quarrying, other mining and quarrying	4744
Provision of services to the mining and quarrying industry and earths	405
Production of food and feed	119677
Beverage production	16080
Tobacco processing	927
Textile production	18692
Clothing production	5863
Production of leather, leather goods and footwear	3060
Manufacture of articles of wood, of straw and of products of wood and cork, except furniture	20092
Manufacture of paper, paperboard and articles thereof	31614
Production of printed matter; reproduction of recorded sound, video and data media	26430
Coking plant and mineral oil processing	2613
Production of chemical products	54767
Production of pharmaceutical products	12703
Manufacture of rubber and plastic products	97564
Manufacture of glass and glass products, ceramics, processing of stones and Earth	50398
Metal production and processing	36651
Manufacture of metal products	175458
Production of data processing equipment, electronic and optical products	56722
Production of electrical equipment	68683
Mechanical Engineering	188739
Manufacture of motor vehicles and parts of motor vehicles	47830
Other vehicle construction	11601
Furniture production	23431
Production of other goods	33746
Repair and installation of machinery and equipment	47480

It can be seen that the industry 2-digit ore mining only has 51 observations in 51 months from 01/2018 to 03/2022. Whether a good model can be learned is currently still questionable. The variables that are the same for all reports are automatically removed as constants by "mice". A growth rate cannot be estimated either, if the only firm has not reported. Also, the time-fixed effects cannot be estimated appropriately in our application. Most industries have at least 4-digit observation numbers. Only three are below that. It is hard to find out how many observations are needed per industry.

B. Relationships between explanatory variables and the variables needed for the Turnover index

18. The coefficient of determination is considered for models in which the target variable is explained by one regressor only.

	Domestic turnover	Foreign turnover	Foreign turnover (non-EU)
Type of Unit	0,0088	0,0073	0,0051
Federal state	0,0049	0,0017	0,0011
Industry 4-digit	0,1215	0,1106	0,0981
Turnover (domestic/foreign/non-EU) of the previous month	0,8992	0,8975	0,8908
Incoming orders (domestic/foreign/non-EU) of the previous month	0,3241	0,7038	0,6798
employees of the previous month	0,2837	0,5915	0,5190

As in the case of new orders, sales are challenging to estimate using the unit type, state, or industry 4-digit only. The coefficient of determination when trying to estimate sales using only the previous month's sales is almost 0.9 for all three types of sales. This value is much higher than estimating new orders using the last month's new orders. This is because order intake fluctuates significantly as a result of major orders. By contrast, turnover is not subject to such fluctuations. The new orders of the previous month are also suitable for estimating the turnover of the current month. The number of employees is relatively constant over time and is also somewhat convenient to explain turnover. It is noticeable that foreign turnover can be better estimated by the new orders of the previous month or the number of employees of the last month than domestic turnover. The reasons for this are still under investigation.

C. Best result to date for the new orders index

19. A variety of models with different regressors were tested. In addition, it was evaluated whether a global model or an individual model for each industry 2-digit performs better. If individual models are learned, this only makes sense on the 2-digit level because there are not sufficient reports per industry on the 3-digit level or even on the 4-digit level.

20. The target variables to be imputed are domestic, foreign, and non-EU new orders. These three target variables are imputed separately. The individual imputation has been established. Otherwise, singularity problems occur. The federal state, the unit type and the industry 4-digit were included in the model because these variables usually have the same value for the current month as in the last month, and thus, information about the current month can be used. Furthermore, the new orders of the previous month and the sales of the previous month were included as Regressors. These variables show a high correlation with the order intake of the current month. Separate models for different order intakes differ in that the domestic order intake is, among other things, explained by the domestic order intake of the previous month. The foreign order intake is explained by its value for the previous month.

The same applies to non-EU new orders and the explanatory variable sales of the previous month. A variable is included in the model that indicates whether a company reported on time or not. This is intended to counteract systematic overestimates or underestimates because, for example, companies with high new orders show different reporting behaviour. It has been shown that it is more useful to learn individual models for the industry 4-digit than a global model for all industries.

There were two versions learned where a global model was learned on the entire data set. These two versions differ in that the average rate of change per industry 2-digit and per reporting period was used as the explanatory variable, and the average rate of change of all observations in this reporting period was used in the second variant. Except for this variable, the model variants are identical. With regard to the MAE and RMSE, the variant with the average rate of change for all industries performed better. For this benchmark model, the MAE was 2.737 and the RMSE was 3.320. For the second model the MAE was 2.834, the RMSE was 3.385, and the POWS was 0.171. However, both models are significantly outperformed when individual models for each industry 2-digit were learned. Here, the MAE was only 2.339, the RMSE was only 3.105, and the POWS was 0,143. This indicates that the assumption of different linear relationships in the various industries is reasonable. The best-performing model is calculated individually for each industry using the following regression:

$$
\begin{aligned}
&\widehat{Incoming}\ oders_{t,i} \\
&\quad = \hat{\beta}_0 + \textbf{Federal state}_{t,i}^{T}\,\hat{\boldsymbol{\beta}}_{\mathbf{1}} + \textbf{Type of unit}_{\mathrm{t,i}}^{\mathrm{T}}\,\hat{\boldsymbol{\beta}}_{2} + \textbf{Industry} - \mathbf{4} - \textbf{digit}_{\mathrm{t,i}}^{\mathrm{T}}\,\hat{\boldsymbol{\beta}}_{3} \\
&\quad + \textbf{Reporting month}_{\mathrm{t,i}}^{\mathrm{T}}\,\hat{\boldsymbol{\beta}}_{4} + \textbf{Turnover class}_{\mathrm{t,i}}^{\mathrm{T}}\,\hat{\boldsymbol{\beta}}_{5} + \textbf{New orders class}_{\mathrm{t,i}}^{\mathrm{T}}\,\hat{\boldsymbol{\beta}}_{6} \\
&\quad + \textbf{Timely notification}_{\mathrm{t,i}}^{\mathrm{T}}\,\hat{\boldsymbol{\beta}}_{7} + \mathrm{Employees}_{\boldsymbol{t-1},i}\,\hat{\beta}_8 + \text{Rate of Chance}_{\boldsymbol{t}}\ \ \hat{\beta}_9
\end{aligned}
$$

For notation: Vectors are bold and scalars are not. Each vector is first a column vector, by transposing we get the row vectors. The bold beta vectors contain, for example, for each state or each type of unit a different parameter value. The vectors as for example federal state contain a 1 in appropriate place and otherwise zeros. t indicates the time and i the i-th operation (Abdalla et al., 2022).

D. Best result to date for the turnover index

21. As in the case of the new orders index, various combinations of explanatory variables were tested for the turnover index. Again, it was tested whether a global model compared to an individual model for each industry performs better.

22. As with the new orders index, the 3 types of sales are estimated separately for the turnover index. The state, the unit type and the industry 4-digit were included in the model because these variables usually have the same value for the current month as in the previous month, and thus, information about the current month can be used. Furthermore, the turnover of the previous month is included as classified variables. This variable has a high correlation with the turnover of the current month. The classification was included to avoid singularity issues. Due to the latter mentioned, the order intake of the previous month was incorporated in the model. However, it is also the case that the previous month's order intake explains about the same portion of the variation as the previous month's sales. Research has shown that a model that only uses the previous month's sales as a regressor has almost the same coefficient of determination as a model that includes both the previous month's sales and the previous month's new orders. As before, the reporting month was included as a variable to reflect time fixed effects. In addition, the growth was incorporated again to take account of the economic situation. The variable

indicating whether a company reports on time is also considered. When individual models were learned for the different industry 2-digits, a MAE of 1.879 and a RMSE of 2.306 were obtained. When the best global model was learned, a MAE of 2.158 and a RMSE of 2.573 is achieved. Only for the POWS do the global models perform better. The POWS is 0.171 for the individual models and 0.143 for the two variants with a global model. Overall, however, the variant with the separate models is still rated as better (Abdalla et al., 2022).

The best-performing model is thus calculated separately for each industry is given by:

$$\widehat{Turnover}_{t,i} = \hat{\beta}_0 + \mathbf{Federal\ state}_{t,i}^{T}\,\widehat{\boldsymbol{\beta}}_1 + \mathbf{Type\ of\ unit}_{\mathrm{t,i}}^{\mathrm{T}}\,\widehat{\boldsymbol{\beta}}_2 + \mathbf{Industry - 4 - digit}_{\mathrm{t,i}}^{\mathrm{T}}\,\widehat{\boldsymbol{\beta}}_3 + \mathbf{Reporting\ month}\,_{\mathrm{t,i}}^{\mathrm{T}}\,\widehat{\boldsymbol{\beta}}_4 + \mathbf{Turnover\ class}\,_{\mathrm{t,i}}^{\mathrm{T}}\,\widehat{\boldsymbol{\beta}}_5 + \mathbf{Timely\ notification}_{\mathrm{t,i}}^{\mathrm{T}}\,\widehat{\boldsymbol{\beta}}_6 + \mathrm{Employees}_{\boldsymbol{t-1},\boldsymbol{i}}\,\hat{\beta}_7 + \mathrm{Rate\ of\ Chance}_{\boldsymbol{t}}\ \ \hat{\beta}_8$$

VI. Conclusion

23. Various imputation methods were tested in this work. Imputation based on a linear regression performed best in the preliminary experiments. It should be noted that much work is ongoing to optimize each method. Good results could already be achieved by some of the methods outlined in the previous chapter, but there is still a lot of optimization potential to improve the results. For the variables that are constant over time but which at least alone do not sufficiently explain the target variable, it is necessary to investigate further how important they are for the model. Furthermore, solutions for the singularity problem have to be found. In addition, new explanatory variables should be investigated, which contain information that the current variables do not yet contain. In addition, it is still possible to set up different models for different sectors of the economy. We conclude that despite potential improvements, regression-based imputation is a useful tool to achieve reasonable estimates for t+15 experimental short-term business statistics.

VII. Bibliography

- Abdalla, S.; Ditscheid, J.; Jäger, J.; Koch, J.; Kremmling, N.; Oruc, B.; Yadegar, E. (2022): „Evaluation des Umsatzindex“
- Abdalla, S.; Ditscheid, J.; Jäger, J.; Koch, J.; Kremmling, N.; Oruc, B.; Yadegar, E. (2022): „Evaluation des Auftragseingangsindex“
- Eurostat (2014): „ESS Handbook for quality reports “
- Hyndman, R.J., und Athanasopoulos, G. (2021): „Forecasting: principles and practice”, 3rd edition, OTexts: Melbourne, Australia. OTexts.com/
- van Buuren S, Groothuis-Oudshoorn K (2011). “mice: Multivariate Imputation by Chained Equations in R.” *Journal of Statistical Software*, **45**(3), 1-67. doi: 10.18637/jss.v045.i03.