LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

INSTITUT FÜR STATISTIK

Christian Seiler & Christian Heumann

Microdata Imputations and Macrodata Implications: Evidence from the Ifo Business Survey

Technical Report Number 119, 2012
Department of Statistics
University of Munich

http://www.stat.uni-muenchen.de

Microdata Imputations and Macrodata Implications: Evidence from the Ifo Business Survey

Christian Seiler*
Ifo Institute

Poschingerstr. 5, 81679 Munich
+49(0)89/9224-1248

Christian Heumann
Department of Statistics,
University of Munich
Ludwigstr. 33, 80539 Munich
+49(0)89/2180-3697

Abstract

A widespread method for now- and forecasting economic macro level parameters such as GDP growth rates are survey-based indicators which contain early information in contrast to official data. But surveys are commonly affected by nonresponding units which can produce biases if these missing values can not be regarded as missing at random. As many papers examined the effect of nonresponse in individual or household surveys, only less is known in the case of business surveys. So, literature leaves a gap on this issue. For this reason, we analyse and impute the missing observations in the Ifo Business Survey, a large business survey in Germany. The most prominent result of this survey is the Ifo Business Climate Index, a leading indicator for the German business cycle. To reflect the underlying latent data generating process, we compare different imputation approaches for longitudinal data. After this, the microdata are aggregated and the results are compared with the original indicators to evaluate their implications on the macro level. Finally, we show that the bias is minimal and ignorable.

JEL Code: C81, C83
Key words: Business survey, Longitudinal data, Imputation, Nonresponse

* corresponding author, Email: seiler@ifo.de

1 Introduction

The usage of survey-based indicators for monitoring as well as now- and forecasting economic parameters has a long tradition in research. For more than 60 years this type of surveys exists and their number has increased throughout the last decades, see Nardo (2003). However, as surveys are commonly affected by nonresponding units, a serious problem may occur when this missing data mechanism includes a selection bias. The evaluation and correction for such biases is especially a concern in household or population surveys and has therefore been discussed extensively in literature. In contrast to this, Janik and Kohaut (2011) mention that only less papers exist with respect to missing data and their effects in business surveys. Therefore, bias patterns may lead to a lower accurancy in forecasting performance of business survey indicators. To fill this gap, we analyse the missing observations in the Ifo Business Survey (IBS), a large monthly business survey in Germany with about 7,000 responding firms each month. Although the IBS has high return rates with more than 85%, nonresponses can cause problems. For example, Schafer (1997) suggests that missing observations are not ignorable when their fraction is higher than 5%. In general, the missing data mechanism is only ignorable if (a) the data are missing at random (MAR)[1] and (b) the parameters for the missing data generating process are unrelated to the parameters the survey is focussed to estimate, see Schunk (2008). Seiler (2010) already showed that the responding behaviour in the

[1]Note that the MAR assumption does not imply that the missing data are a random subset of the entire data set. The latter is called 'missing completely at random' (MCAR) and is even more restrictive. See Little and Rubin (2002) for definitions.

IBS depends to a minor extend on the business cycle, i.e. the interesting latent variable, after controlling for survey-related effects. These findings suggest that assumption (b) could be violated.

For this reason, we develop different imputation strategies for the missing microdata sets in the IBS and analyse their effects on the aggregated macro indicators, i.e. the Ifo Business Climate Index, a leading indicator for the German business cycle. After imputation, we are able to investigate the presence and, if existing, the magnitude of a possible bias also with respect to forecasting issues. Although a general problem in studies regarding imputation analysis prevails that the MAR or MCAR assumption can not be tested when there exists no additional information about the data (see Manski (2003) and Cameron and Trivedi, 2005), we presume to find an appropriate imputation model with reflects the inherent dynamics and leads to good estimates since the survey has high frequency and the interesting variables change relatively smoothly over time. So, conditionally on our model we evaluate from our data, we assume that MAR is fulfilled.

Therefore, the paper is organised as follows: The data set and its specifics are discribed in Section 2. We show some descriptive statistics and display how the survey is performed and structured according to EU regulations. In Section 3 we develop different imputation strategies for these specific kind of data and compare the power of these imputation approaches. Section 4 shows and compares the aggregated results after imputation of missing values. We analyse these macrodata time series up to results for the sublevels and finally compare the forecasting performance of the original and imputed Business Climate Index. Section 5 sums up our empirical findings.

2 Data

2.1 The Survey

The development of survey-based business cycle indicators has its seeds in the need of early information on the economic development. As official data are published with high delay and also commonly revised after the first publication, business cycle tendency surveys can considerable quicker monitor the actual economic situation. The Ifo Institute was one of the first when conducting its *Ifo Business Survey* in 1949 and within the last 60 years this method has been accepted widely in the OECD countries, see OECD (2003) and Nardo (2003). In line with the Joint Harmonised EU Programme of Business and Consumer Surveys (see European Union, 2006), these indicators base on two variables (business situation and business expectations) which are measured on a 3-level-Likert scale representing a good, equal or bad state. Due to the construction of the questions in the questionnaire, the resulting indicators in fact measure the business cycle without trend (OECD, 2003).

The data used in this paper are from the Ifo Business Survey, the German part of the Joint Harmonised EU Programme. The most well-known result of this survey is the *Ifo Business Climate Index*, a leading indicator for the German business development which is used for forecasting analyses. Every month about 7,000 companies respond.[2] For further methodological information on this survey see Goldrian (2007) and the early works of An-

[2]The data sets are available at the Economics & Business Data Center (EBDC), a combined platform for empirical research in business administration and economics of the Ludwig Maximilian University of Munich (LMU) and the Ifo Institute.

derson (1951, 1952) and Theil (1952). Becker and Wohlrabe (2008) give an overview on the collected variables and Abberger and Wohlrabe (2006) on the literature with respect to forecasting analyses with the Ifo index.

As stated above, the Ifo index is constructed using only two variables of the survey: The actual business situation (BS) in the appropriate month and the business expectations (BE). Both are measured on a 3-level Likert-scale with values 'good'/'better' (indicated by +), 'satisfactory'/'about the same' (=) or 'bad'/'worse' (-). To calculate the index, the answers are weighted by the companies' size (which is updated once a year) and the area the firm works according to the official classification from the German Statistical Office.[3] To achieve the final value of the index, the fraction of negative replies is substracted from the positive ones (for each variable) in a first step and then the harmonic mean is taken to construct the 'business climate' from the 'business situation' and 'business expectation' values. The aggregation scheme is presented in appendix A. In principle, the index can be interpreted as some kind of a weighted mean. Due to about 7,000 respondents every month, also indicators for lower aggregation levels are calculated.

Since more than 99.9% of missing values for BS and BE are due to unit-nonresponse, we do not perform an analysis by imputing only item- but not unit-nonresponse because we do not expect any major differences to the original results with missing data. Therefore, no other variables from the survey (such as demand or production) could be used as covariates, as we have no answers from the corresponding company at the time of unit-

[3]The second is done due to the fact that the survey is not a random sample of all companies in Germany.

nonresponse. However, in Section 3 we show how these strong assumptions can be relaxed so that regression approaches are possible. In addition, we make use of the dynamics of the panel structure and additionally are able to examine similarity relations based on the characteristics of the firms.

2.2 Some descriptive statistics

To get an idea of the extent of the variables and the missing values in the IBS, we provide some descriptive statistics in this Subsection. Our variables of interest are both measured on a 3-level Likert scale, so every company changes over time between these three states and nonresponse can be treated as a fourth state. Table 1 shows the stochastic matrices for BS and BE for the six transition periods $t-1 \rightarrow t, \ldots, t-6 \rightarrow t$. The probability for staying in the same state is relatively high, so the state change is slow in relation to the survey frequency. It can also be seen that the business expectations change more often than the business situation which is rather unsurprising. A striking fact is that the probability for changing from responding to nonresponding from t to $t-1$ is different for the state of the company conditional on $t-1$. The probability not to respond in period t is almost twice as high after responding 'bad' in contrast to 'good' in $t-1$. This is in contrast to macro level results, where nonresponses are more frequent in economic better times, see Harris-Kojetin and Tucker (1999) and Seiler (2010). In addition, switching from nonresponse to response also seems to be selective since the probabilities of the categories are not equal. If the firms leave the nonresponse-'state', e.g. $P(x_t \equiv + | x_{t-1} \equiv NA)$, only 9% of the firms replied

a positive business situation. According to Little and Rubin (2002), MAR is not fulfilled in panel data sets if the probability for missing depends on the *future* values of x which suggests that the data could be biased. Since we

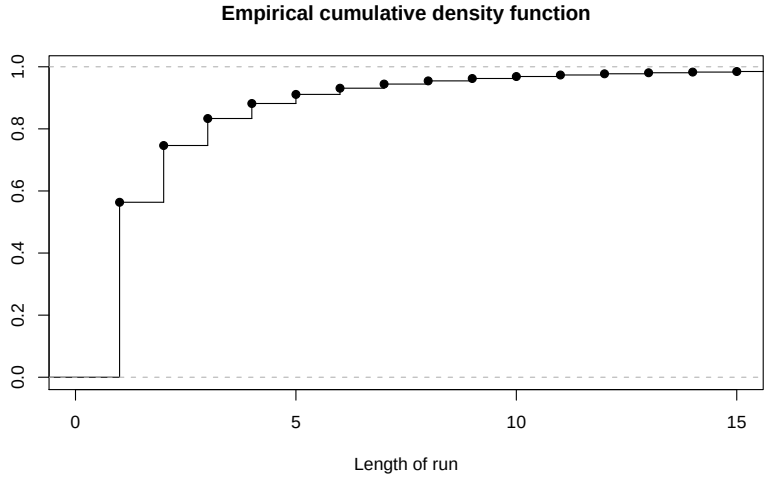

Figure 1: Empirical cumulative density function for the length of successive unit nonresponse

have repeated measures from the same units, we have to get an idea of the presence of sequences of missings in our data. Figure 1 shows the length of successive unit nonresponse. 56% of the missing data is due to nonresponse only for a single month. More than 80% of nonresponse appears within 3 months and 90% within 6 months. Thus, for most values prevailing information is still available. Depending on the imputation method (especially conditional models), probably not all missing values can or should be estiamted as one can expect that the predictive power of the model decreases when many successive missings occur. Therefore, we validate our results in Section 4 according to different horizons h of successive nonresponse.

$t-1/t$	+	=	-	NA
+	0.68	0.24	0.03	0.05
=	0.06	0.75	0.12	0.07
-	0.01	0.15	0.75	0.09
NA	0.04	0.20	0.20	0.56

$t-1/t$	+	=	-	NA
+	0.58	0.32	0.04	0.06
=	0.07	0.76	0.11	0.07
-	0.02	0.27	0.62	0.09
NA	0.05	0.26	0.12	0.56

$t-2/t$	+	=	-	NA
+	0.62	0.29	0.04	0.05
=	0.07	0.71	0.15	0.07
-	0.01	0.19	0.71	0.09
NA	0.04	0.22	0.21	0.52

$t-2/t$	+	=	-	NA
+	0.50	0.38	0.05	0.07
=	0.08	0.72	0.13	0.08
-	0.04	0.31	0.56	0.09
NA	0.06	0.29	0.13	0.52

$t-3/t$	+	=	-	NA
+	0.58	0.32	0.05	0.05
=	0.07	0.68	0.17	0.07
-	0.02	0.21	0.68	0.09
NA	0.05	0.24	0.22	0.49

$t-3/t$	+	=	-	NA
+	0.45	0.41	0.07	0.07
=	0.08	0.70	0.14	0.08
-	0.05	0.34	0.52	0.09
NA	0.07	0.31	0.14	0.49

$t-4/t$	+	=	-	NA
+	0.54	0.34	0.06	0.05
=	0.08	0.66	0.18	0.08
-	0.02	0.22	0.66	0.10
NA	0.05	0.26	0.23	0.46

$t-4/t$	+	=	-	NA
+	0.41	0.43	0.09	0.07
=	0.08	0.69	0.15	0.08
-	0.06	0.36	0.49	0.10
NA	0.07	0.32	0.14	0.46

$t-5/t$	+	=	-	NA
+	0.52	0.35	0.07	0.06
=	0.08	0.65	0.19	0.08
-	0.02	0.23	0.65	0.10
NA	0.06	0.27	0.23	0.44

$t-5/t$	+	=	-	NA
+	0.38	0.44	0.10	0.07
=	0.09	0.68	0.15	0.08
-	0.07	0.37	0.47	0.10
NA	0.07	0.34	0.14	0.44

$t-6/t$	+	=	-	NA
+	0.49	0.36	0.09	0.05
=	0.08	0.64	0.20	0.08
-	0.02	0.24	0.63	0.10
NA	0.06	0.28	0.24	0.42

$t-6/t$	+	=	-	NA
+	0.36	0.45	0.12	0.07
=	0.09	0.67	0.16	0.08
-	0.07	0.38	0.45	0.10
NA	0.08	0.35	0.15	0.42

Table 1: Stochastic matrices for business situation (left) and business expectation (right) for $t-1 \to t, t-2 \to t, \ldots, t-6 \to t$ (top to bottom)

3 Methodology

3.1 Requirements on the imputation methods

The methodologic beginnings of imputation models for missing observations are mostly associated with Donald Rubin, see in particular Rubin (1987) and Little and Rubin (2002) for an overview and definition of missing data patterns. In literature, a wide variety of different imputation methods exist but the data in this paper has special structure so every imputation approach has to face strong requirements. As noted in Section 2 our variables of interest are measured on a 3-level Likert scale. In principle, both variables are expressions of a latent variable (the 'business situation' and the 'business expectations') which is supposed to change over time depending on the business cycle. This fact implies two requirements on the imputation methods: First, as we analyse panel data, we have to use imputation methods which can reflect the inherent dynamics of the underlying latent process. This means that t, the calendar time, should be included in some form into the imputation model. Engels and Diehr (2003) and Kleinke et al. (2011) give an overview on the imputation of panel data but also mention that most standard approaches implemented in statistical software packages are limited to handle incomplete panel data. Second, as our variables of interest have only three different states, we have to choose methods which impute plausible values. For this reason, many approaches such as simple mean imputation can not be used here as they require a continuous variable to impute, see Finch (2010) for an overview. Based on this structure we con-

sider our data as Markov chains for every unit with time-inhomogenous transition matrices. Therefore, also imputation approaches developed for time-series data but with continuous range are not an appropriate solution here.

In addition to the requirements stated above, we have to face some other issues in our analysis. Major problem of the data set is the fact that missing observations appear almost solely due to unit nonresponse. For this reason, there are hardly any covariates at the same point in time when nonresponse occurs, so that a regression analysis (or more general: external information) can, in principle, not be used. This would reduce the number of eligible imputation methods enormously, but we will later show how these strong assumptions can be relaxed. Basically, two general approaches to include explanatory information remain: On the one hand, using the individual past and their inherent dynamics. On the other hand, using attributes from similar companies at the same time after defining a simliarity structure. Another problem is the extent of the data set when running multiple imputations. Since we have more than 1.6 million observations, multiple imputation (MI) would increase computing time. Graham et al. (2007) argue that the number of imputations should be higher than usually expected, but they also notice that this depends on the researchers tolerance of precision and the computing time to run these multiple imputations. The most common choice in literature is to set the number of multiple imputations to 5 which is also done here. In general, MI is only appropriate for probabilistic approaches, because deterministic methods anyway only lead to a single value. So the strategy is as follows: We try to find the imputation method

which reflects the data generating process best and use this method to impute the missing values. To decide between the approaches, we introduce a measure to evaluate the predictive power in Section 3.3.

3.2 Imputation methods for ordinal panel data

3.2.1 Last observation carried forward

One of the easiest ways to impute longitudinal data is the *last observation carried forward* (LOCF) method. The idea behind is quite simple: If nonresponse occurs, the last recorded observation of the interesting unit is taken. So, this method needs the strong assumption that the value remains unchanged in case of nonresponse. Little and Rubin (2002) argue that this assumption is unrealistic in many settings. In recent years, this approach came more and more under criticism, see for example Cook et al. (2004) and Saha and Jones (2009). Nevertheless, LOCF is widely used, particularly in clinical studies (see Woolley et al., 2009). We assume that this method leads to relatively good results in cases of ordinal data with less states and if the number of successive missing values is not too long. For our data, both arguments seem to be the case. However, from Section 2.2 we know that long runs of missings are seldom but can occur. Therefore, we have to evaluate the power of this method according to the length of successive missing values, because it is plausible that predictive power decreases if the last recorded observation dates back several months. Due to its intensive use of LOCF, it is also a good proxy for other imputation methods. We expect that a structured approach should be able to produce better estimates than LOCF.

3.2.2 Nearest neighbour

The *nearest neighbour* method (NN) is a very wide class of imputation approaches and is also one of the most commonly used. The basic idea behind is to find a 'donor' for the 'recipient', i.e. an observation with same or similar properties and a recorded value which is then transfered to the nonresponding unit. Chen and Shao (2000) give an overview on the consistency of NN imputation. This approach can be extended to draw from a distribution or take the mode if more than one 'donor' is available which is known as k-NN imputation. In our setting, we assume similarities according to the same business area in the appropriate month. This means that the number of possible donors k differs between sectors and survey waves. The main reason why k has to be flexible in our case is that there are no other sensible variables which can be used to define a similarity structure. As only one variable (the business area) remains, we use all possible units for this approach. So, our imputation strategy is as follows: For every month, we calculate the distribution for the three states (+, =, -) according to a specific business area. For the nonresponding firms, we draw from this distribution.[4] This means we assume that missing units behave as the observed companies from the same business area.

3.2.3 Markov Chains

As can be seen in Table 1, the probability for staying in the same state is relatively high even after six months. In order to use this fact, we consider in-

[4]The mode is not used in this case, because at almost every time the proportion of '='-responses is largest and hence we would impute only '='-values.

dividual state changes as a *Markov Chain* (MC). Therefore, let $X_i = \{X_{i,t}, t \in T\}, i = 1, \ldots, n$, be a stochastic process for every unit i representing BS or BE by a given probability space $(\Omega, \mathcal{F}, \mathcal{P})$. In our case, we are interested in calculating the stochastic matrices

$$P = (p_{rs}), \quad r, s \in S,$$

$p_{rs} = P(X_t = s | X_{t-1} = r)$ and $S = \{+, =, -\}$. Because of our presumption in Section 3.1 that there exists an underlying dynamic process, we assume that the stochastic matrix P is time-dependent, so

$$P = P(t) = (p_{rs}(t))$$

which means that X is an inhomogeneous Markov Chain. In Table 1 we showed that the probabilities for staying in the same state are quite high, so that this method would not make any difference to LOCF if we take the mode and we assume that the highest probabilities are on the main diagonal for every t. For this reason, we take a step beyond and extend the Markov Chains to order k. So, the stochastic matrix is

$$P^k(t) = (p_{(r_{t-1},\ldots,r_{t-k},s)}(t)),$$

$p_{(r_{t-1},\ldots,r_{t-k},s)}(t) = P(X_t = s | X_t = r_{t-1}, \ldots, X_{t-k} = r_{t-k}, t)$. Notice that $\dim(P^k(t)) = |S|^k \times |S|$, so when k increases by 1 the number of rows of $P^k(t)$ increase by the factor $|S| = 3$. Less technically speaking, this means that we evaluate the runs of answers of the last k months and calculate the

probabilities for different states in t. This procedure is done for every t, so we produce 'rolling' stochastic matrices. We hope that with this method a good classification of the companies can be obtained and we receive high probabilities for at least one of the three states in every row of $P^k(t)$. After evaluating the stochastic matrices $P^k(t)$, every company with missing data in t is classified by their past k values and finally we draw from this distribution or take the mode, i.e. the state with the highest probability in t, to impute the missing value. The higher k is set, the higher the specialisation is but the more transitions have to be evaluated. For $k = 5$ there are $3^5 = 243$ transitions. In spite of the large data set, many transitions do not occur in the data and therefore we set the maximum for k to 4, i.e. 81 possible transitions. We notice that this approach is uncommon in imputation analysis but results from data structure and effectively is the equivalent to an AR process on macro level which plays a major role in every forecast analysis of time series. In fact, this approach is the same as a nearest neighbour imputation with a smiliarity structure defined on the past k months.

3.2.4 Joint distribution

The assumptions for the MC approach are relatively restrictive as the firms do have to have the exact transition of their answers to be a possible candidate for imputation. A more flexible method would be to focus on the *joint distribution* (JD) of BS and BE and the individual past $t-1, \ldots, t-k$ of both variables, i.e.

$$f_{JD,k}(t) := f(BS_t, BS_{t-1}, \ldots, BS_{t-k}, BE_t, BE_{t-1}, \ldots, BE_{t-k}, t). \tag{1}$$

The most frequent approach to obtain such joint probability functions $f_{JD,k}(t)$ in order to impute missing data is done with the `Amelia` package, version 1.5-4 developed by Honaker et al. (2011) and originally proposed in King et al. (2001). `Amelia` requires that the joint distribution in equation (1) is multivariate normal, which is obviously violated in our case. Fortunately, the `Amelia` package also provides imputation of ordinal variables by transformation. As we analyse panel data, `Amelia` enalbes to specify time and cross sectional variables. In addition, time-varying effects as well as lags of the interesting variables can be included into the imputation model, see Honaker and King (2010). Therefore, this approach is very flexible and can reflect both, the individual state change as well as the overall underlying latent process.

3.2.5 Regression approaches

All approaches mentioned above are relatively easy to implement but do not inherent economic relationships. As regression models could not be used to due to the fact that no information is avialable at the same time because of massive unit-nonresponse, we notice that no 'real' explanatory variables (ignoring the case that the business area and the individual past can in some sense also be regarded as explanatory information) are on hand. In this Subsection, we will show how these strong assumptions can be relaxed.

Due to the ordinal structure of our variables of interest, a regression-based imputation approach would have form of a *proportional odds model*

(McCullagh, 1980)

$$\eta_i = g(\mu_i) = \log \frac{P(y_i \leq c|\boldsymbol{x}_i)}{P(y_i > c|\boldsymbol{x}_i)} = \tau_c - \boldsymbol{x}_i\boldsymbol{\beta}, \qquad c = 2, \ldots, C, \tag{2}$$

where

$$\mu_i = \mathrm{E}(y_i|x_i) = h(\eta_i), \qquad h(\cdot) = g^{-1}(\cdot)$$

and $C = 3$ as both variables of interest are measured on a 3-level Likert scale. Model (2) has the advantage that it models a latent variable and calculates thresholds τ_c. As noted in Section 2.1, due to unit nonresponse no covariates are available. But from Section 2.2 we know that the companies remain relatively long in the same state and that this change is slow in relation to the survey frequency. Now it is assumed that this applies also for the other variables of the survey, which are also mainly be measured on a 3-level-Likert scale. So, besides the individual past, $\boldsymbol{x}_i = \boldsymbol{x}_{i,t-1}$ contains additional variables asked in the survey from the preceeding month.[5] As the individual past of the depending variable BS or BE is included into $\boldsymbol{x}_{i,t-1}$, model (2) enables to check whether the inclusion of additional explanatory variables improves the estimation of BS and BE.

The major disadvantage of this approach is the fact that model (2) can only be estimated when all variables are observed. In cases of two or more successive unit nonresponse $\boldsymbol{x}_{i,t-1}$ is missing, i.e. $\boldsymbol{x}_{i,t-1}$ itself has to be imputed. This exacerbates the problem since we have to find an appropriate

[5]Considering Section 1, we interpret wave $t-1$ as a 'representative' for wave t due to the high frequency of the survey.

model for every variable in $\boldsymbol{x}_{i,t-1}$, which is not done in this paper. The analyses are therefore restricted to impute only months after the firm responded, so we are only able to impute at least 56% of our missing data. Another issue is that the questions[6] on the questionnaire differ highly between the sectors. For example, the degree of capacity utilisation is asked in construction, but obviously not in trade. For this reason, we calculate a different model of form (2) for each of the three sectors with sector-specific covariates $\boldsymbol{x}_{t-1}^{sec}$. Table 4 provides an overview. In addition, we need to evaluate different models depending on t to reflect the inherent dynamics. Thus, a separate model

$$\eta_{c,t}^{sec} = \log \frac{P(y_t \leq c|\boldsymbol{x}_{t-1}^{sec})}{P(y_t > c|\boldsymbol{x}_{t-1}^{sec})} = \tau_{c,t}^{sec} - \boldsymbol{x}_{t-1}^{sec}\boldsymbol{\beta}_{c,t}^{sec}$$

with $t = 1, \ldots, 192, c = 2, 3$, for each t is calculated.

3.3 Goodness of fit

To decide which method explains the data best, we have to introduce a statistical measure to evaluate the goodness of fit for the estimators of the different imputation methods. As our variables of interest are discrete, we can count the number of correct and incorrect predicted values in a 3×3-matrix. Therefore, we introduce *Cohen's kappa* (Cohen, 1960) which is defined as

$$\kappa = \frac{\pi_o - \pi_e}{1 - \pi_e},$$

[6]Moreover, all variables included in $\boldsymbol{x}_{i,t-1}$ are restricted to those who are measured monthly.

where $\pi_o = \sum_{c=1}^{C} \pi_{cc}, C = 3$, is the relative observed correspondence of the estimators, and $\pi_e = \sum_{c=1}^{C} \pi_{c\cdot}\pi_{\cdot c}$ the hypothetical probability of correspondence when there is no relationship between the original and imputed values. There also exists a weighted version of Cohen's kappa with weights w_{cd}, leading to $\pi_o = \sum_{c=1}^{C}\sum_{d=1}^{C} w_{cd}\pi_{cd}$ and $\pi_e = \sum_{c=1}^{C}\sum_{d=1}^{C} w_{cd}\pi_{c\cdot}\pi_{\cdot d}$. If you use, for example, quadratic weights

$$w_{cd} = 1 - \frac{(c-d)^2}{(l-1)^2} = 1 - \frac{(c-d)^2}{4},$$

these would give a weight of 1 to the diagonal elements, i.e. the correct imputed values, 0.75 to the adjacent categories (to 2 if 1 or 3 is correct and to 1 and to 3 if 2 is correct) and 0 in the other cases. In this paper, the unweighted version of κ is calculated. This is more restrictive than the weighted version but since there are only 3 different possible states the imputation method should be good enough to estimate the real value, in particular as only 2 out of 9 combinations would have an weight of 0. To calculate Cohens kappa for all observed units we make use of the leaving-one-out principle, i.e. we treat every observed unit i as missing and construct the imputation method based on this reduced data set. This approach de facto leads to an overimputation of the whole data set.[7]

For $\kappa > 0$, the estimator provides an improvement over a pure random estimate. Note that the theoretical maximum of 1 is only reached when row

[7]However, for JD, we have to adapt this principle. As JD calculates time and lag effects in closed form for the whole data set, leaving-one-out would increase computing time enormously as 1.6 million observations enter the imputation model and beyond this has to be done for all of these 1.6 million observations. To calculate Cohens kappas for this approach, we randomly drop about 20% of our observed data and tests the power of this imputation method with these values.

and column sums are identical. In all other cases, $\max(\kappa)$ is smaller than 1. Due to large number of observations in our data the maximum is actually nearly 1. Since two variables are imputed, we calculate Cohen's kappa for both, BS and BE separately.

In Section 2.2, Figure 1, we showed that about 60% of our missing values are missings after at least one missing occured in the previous wave. If we use an imputation approach depending on the individual past (i.e. $f(y_t|y_{t-1},\ldots,y_{t-k})$) and an unit which has more than one missing in a row, we are then able to impute the first missing as usual. But if the second missing in a row is about to be imputed, we would depend this method on an estimated value, i.e. $f(y_{t+1}|\hat{y}_t, y_{t-1},\ldots,y_{t-k+1})$, which might cause more uncertainty. Therefore, we have to check how the imputation method works on longer runs of missings. This is done by calculating κ's for a 'forecasting horizon' h of up to 6 months. In Section 4, we also calculate the indicators based on an imputation for different lengths of successive missings, i.e. for $h = 1, 3, 6$ and $\max(h)$, to display the differences.

3.4 Comparison

Tables 2 and 3 show Cohen's kappas for BS and BE and the different imputation approaches. To assess the strength of imputation Landis and Koch (1977) introduced the following rule of thumb: $\kappa < 0$ indicates no agreement, for $0 \leq \kappa \leq 0.2$ agreement is slight, $0.2 < \kappa \leq 0.4$ is fair, $0.4 < \kappa \leq 0.6$ is moderate, $0.6 < \kappa \leq 0.8$ is substantial, and $0.8 < \kappa \leq 1$ is almost perfect. First, notice that κ_{BE} is always smaller than κ_{BS}. This result is not surprising

since expectations are directed to the future and therefore more difficult to assess than the present situation. It is striking that LOCF, which is a rather simple method, produces relatively good estimates. Up to six months the agreement is still moderate for BS and fair for BE. It is also not surprising that the quality of imputation becomes worse as more consecutive missing values have to be imputed. Note that for the Markov Chain approaches, $\kappa_h = \kappa_{k+1}$ if $h \geq k+1$ because after k months these approaches depend only on estimated values.

In contrast to LOCF, the nearest neighbour approach clearly performs worse. Since imputations are drawn from the populations' distribution, we only replicate these probabilities. As no past information enters this method, all κ's are equal regardless of horizon h. Therefore, this method is only slightly better than randomisation, since specialisation by business area seems to contain only minor information. A higher specialisation is possible, but this would increase computing time enormously and it is unlikely that higher values of kappa comparing to LOCF can be obtained. All proportional odds models perform relatively good but on average they are not better than LOCF. In addition, all of these models are restricted to an imputation of the first month a missing value occurs. A good performance (for BS) is obtained when using the Markov Chain approach and taking the mode.[8] In general, drawing from the distribution of the calculated stochastic matrices is always worse than taking the mode. For BS, the best result can be achieved by using the modes of the Markov Chains of order 2. Although

[8] As the fraction of 'equal' answers is the highest in nearly all of the months, we did not calculate MC1 (M) because this is equal to LOCF.

LOCF performs slightly better compared to MC2 (M) for $h = 2, 3$ and 4, MC2 (M) has the big advantage that for $h \geq 3$ it imputes 67% of our data correct although it only depends on estimated values. However, a higher differentiation by including more months into the stochastic matrices does not lead to a better imputation performance. Therefore, only MC2 (M) provides an improvement over LOCF as it seems to reflect the current dynamics better. The joint distribution evaluated by the `Amelia` package seems to include even more uncertainty as the MCk (D) approaches and also performs worse. For BE, none of the other approaches beats LOCF. For this variable, it seems to be hard to find a real structured model. However, the reader should keep in mind that imputation models are predictive and not causal (Honaker et al.,2011), so every imputation approach is only measured by its estimation performance.

Abbr.	Method	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$	$h=6$
JD	Joint Distribution	0.222 (48%)	0.008 (34%)	0.007 (34%)	0.011 (34%)	0.001 (33%)	0.010 (34%)
LOCF	Last observation carried forward	0.668 (78%)	0.590 (73%)	0.542 (69%)	0.504 (67%)	0.476 (65%)	0.452 (63%)
MC1 (D)	Markov Chains (last month, distribution)	0.492 (66%)	0.238 (49%)	0.238 (49%)	0.238 (49%)	0.238 (49%)	0.238 (49%)
MC2 (D)	Markov Chains (last 2 months, distribution)	0.470 (65%)	0.344 (56%)	0.179 (45%)	0.179 (45%)	0.179 (45%)	0.179 (45%)
MC3 (D)	Markov Chains (last 3 months, distribution)	0.489 (66%)	0.389 (59%)	0.008 (34%)	0.004 (34%)	0.004 (34%)	0.004 (34%)
MC4 (D)	Markov Chains (last 4 months, distribution)	0.044 (36%)	0.035 (36%)	0.036 (36%)	0.035 (36%)	0.033 (36%)	0.033 (36%)
MC2 (M)	Markov Chains (last 2 months, mode)	0.674 (78%)	0.570 (71%)	0.501 (67%)	0.501 (67%)	0.501 (67%)	0.501 (67%)
MC3 (M)	Markov Chains (last 3 months, mode)	0.655 (77%)	0.581 (72%)	0.240 (49%)	0.212 (47%)	0.212 (47%)	0.212 (47%)
MC4 (M)	Markov Chains (last 4 months, mode)	0.391 (59%)	0.348 (57%)	0.307 (54%)	0.256 (50%)	0.219 (48%)	0.219 (48%)
NN-II	Nearest Neighbour (level II)	0.066 (38%)	0.066 (38%)	0.066 (38%)	0.066 (38%)	0.066 (38%)	0.066 (38%)
NN-III	Nearest Neighbour (level III)	0.093 (40%)	0.093 (40%)	0.093 (40%)	0.093 (40%)	0.093 (40%)	0.093 (40%)
POM-IND	Proportional odds model (industry)	0.640 (76%)	-	-	-	-	-
POM-CON	Proportional odds model (construction)	0.700 (84%)	-	-	-	-	-
POM-TRA	Proportional odds model (trade)	0.552 (72%)	-	-	-	-	-

Table 2: Overview of imputation methods for variable *business situation* (BS): Cohens kappas and the fraction of correct imputed values (in brackets under the values of κ) by horizon h. For Markov Chains approaches, (M) denotes usage of the mode whereas (D) denotes drawing from the calculated distribution. κ's for all probabilistic methods (JD, MC (D) and NN) are calculated by the average of 5 replications.

Abbr.	Method	$h=1$	$h=2$	$h=3$	$h=4$	$h=5$	$h=6$
JD	Joint Distribution	0.122 (41%)	0.003 (34%)	0.003 (34%)	0.007 (34%)	0.001 (33%)	0.013 (34%)
LOCF	Last observation carried forward	0.544 (70%)	0.435 (62%)	0.398 (60%)	0.358 (57%)	0.329 (55%)	0.308 (54%)
MC1 (D)	Markov Chains (last month, distribution)	0.295 (53%)	0.150 (43%)	0.150 (43%)	0.150 (43%)	0.150 (43%)	0.150 (43%)
MC2 (D)	Markov Chains (last 2 months, distribution)	0.118 (41%)	0.087 (39%)	0.033 (36%)	0.033 (36%)	0.033 (36%)	0.033 (36%)
MC3 (D)	Markov Chains (last 3 months, distribution)	0.124 (42%)	0.102 (40%)	0.006 (34%)	0.000 (33%)	0.000 (33%)	0.000 (33%)
MC4 (D)	Markov Chains (last 4 months, distribution)	0.013 (34%)	0.008 (34%)	0.010 (34%)	0.011 (34%)	0.010 (34%)	0.010 (34%)
MC2 (M)	Markov Chains (last 2 months, mode)	0.173 (45%)	0.155 (44%)	0.038 (36%)	0.038 (36%)	0.038 (36%)	0.038 (36%)
MC3 (M)	Markov Chains (last 3 months, mode)	0.172 (45%)	0.158 (44%)	0.062 (37%)	0.038 (36%)	0.038 (36%)	0.038 (36%)
MC4 (M)	Markov Chains (last 4 months, mode)	0.098 (40%)	0.089 (39%)	0.077 (38%)	0.051 (37%)	0.021 (35%)	0.021 (35%)
NN-II	Nearest Neighbour (level II)	0.035 (36%)	0.035 (36%)	0.035 (36%)	0.035 (36%)	0.035 (36%)	0.035 (36%)
NN-III	Nearest Neighbour (level III)	0.056 (37%)	0.056 (37%)	0.056 (37%)	0.056 (37%)	0.056 (37%)	0.056 (37%)
POM-IND	Proportional odds model (industry)	0.464 (66%)	-	-	-	-	-
POM-CON	Proportional odds model (construction)	0.320 (59%)	-	-	-	-	-
POM-TRA	Proportional odds model (trade)	0.310 (58%)	-	-	-	-	-

Table 3: Overview of imputation methods for variable *business expectations* (BE): Cohens kappas and the fraction of correct imputed values (in brackets under the values of κ) by horizon h. For Markov Chain approaches, (M) denotes usage of the mode whereas (D) denotes drawing from the calculated distribution. κ's for all probabilistic methods (JD, MC (D) and NN) are calculated by the average of 5 replications.

4 Bias analysis

4.1 Visual inspection

After imputing the missing values according to MC2 (M) (for BS) and LOCF (for BE) and running the aggregation scheme displayed in appendix A, we are able to compare the indices with imputed missing values with the original ones. We also run imputations for four different horizons h, as mentioned in Section 3.3. For level 0, the indices for business situation, business expectations and the composed business climate are displayed in Figure 2.

It can easily be seen that the difference between both indices is small. The maximum difference is about 0.02 which is very low in comparison to the indicators' range.[9] As we can not display all time series for all of the sublevels, we draw boxplots for the distribution of the absolute differences according to level and horizon in Figure 3. In general, the absolute differences increase with the level. This is not surprising as the number of observations get lower and the imputed firms obtain more weight in the subgroups' indicators. Nevertheless, the maximum difference found in our data is around 0.15. Also, the difference rises with a higher horizon h. As more missing values are imputed, the average difference between two indicators increases. However, as we also imputed up to $h = \max(h)$, we can seen that the average difference does not rise too strongly compared to h.

Even if the absolute difference is small, Figure 2 shows that the differences seem to depend from the underlying variable, i.e. the bias, and therefore the missing observations, seem not to be random. To check this as-

[9] The theoretical range is $[-1, 1]$.

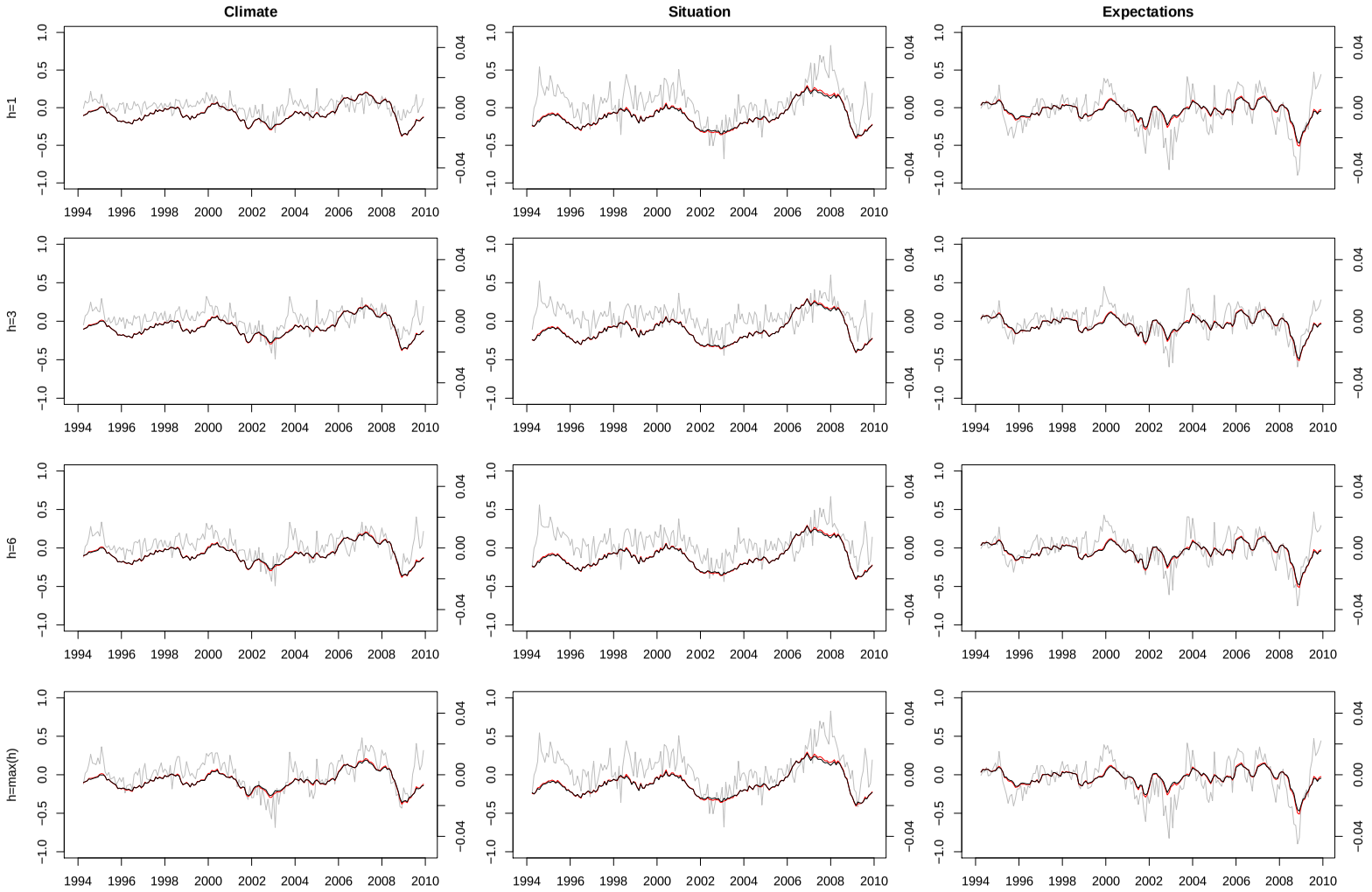

Figure 2: Original (black) and imputed (red) Ifo indicators and their difference (gray, right scale)

sumption and to evaluate the magnitude of the dependence from the underlying variable, we calculate Spearmans correlation coefficient ρ_{GDP} between the imputed indicator and the growth rates of the German Gross Domestic Product, which are the most common 'expression' of the business cycle. For level 0 and $h = 1$, $\rho^0_{GDP} = 0.452$, which means that the bias is relatively strongly correlated with the business cycle. Figure 4 shows the boxplots according to level and horizon h. With increasing h, the correlations ρ_{GDP} do not seem to become higher. However, the correlations decrease with the levels, but this may due to a lower dependence from the business cycle in the sublevels. Because it is hard to find a time series for every business area which reflects the business cycle in this area at best, we also calculate the correlations between the differences and the imputed indicators ρ_{IND}. Figure 5 shows their distributions. The correlations are higher than for the correlations with the GDP and rise, on average, with horizon h. In general, the visual inspection shows that the bias is minimal but related to the underlying variable. To a very small extent upper turning points are underestimated whereas lower turning points are overestimated.

However, Figure 2 also suggests that the indicators are strechted due to the nonresponse bias. These effects may for example occur when the 'equal'-category in underrepresented. As these indicators are artificial by definition, such a stretch would not lead to an substantial change in interpretation as the absolute value of the indicator does not reflect a certain quantity (e.g. in contrast to the GDP). Therefore, we standardise all indicators (original as

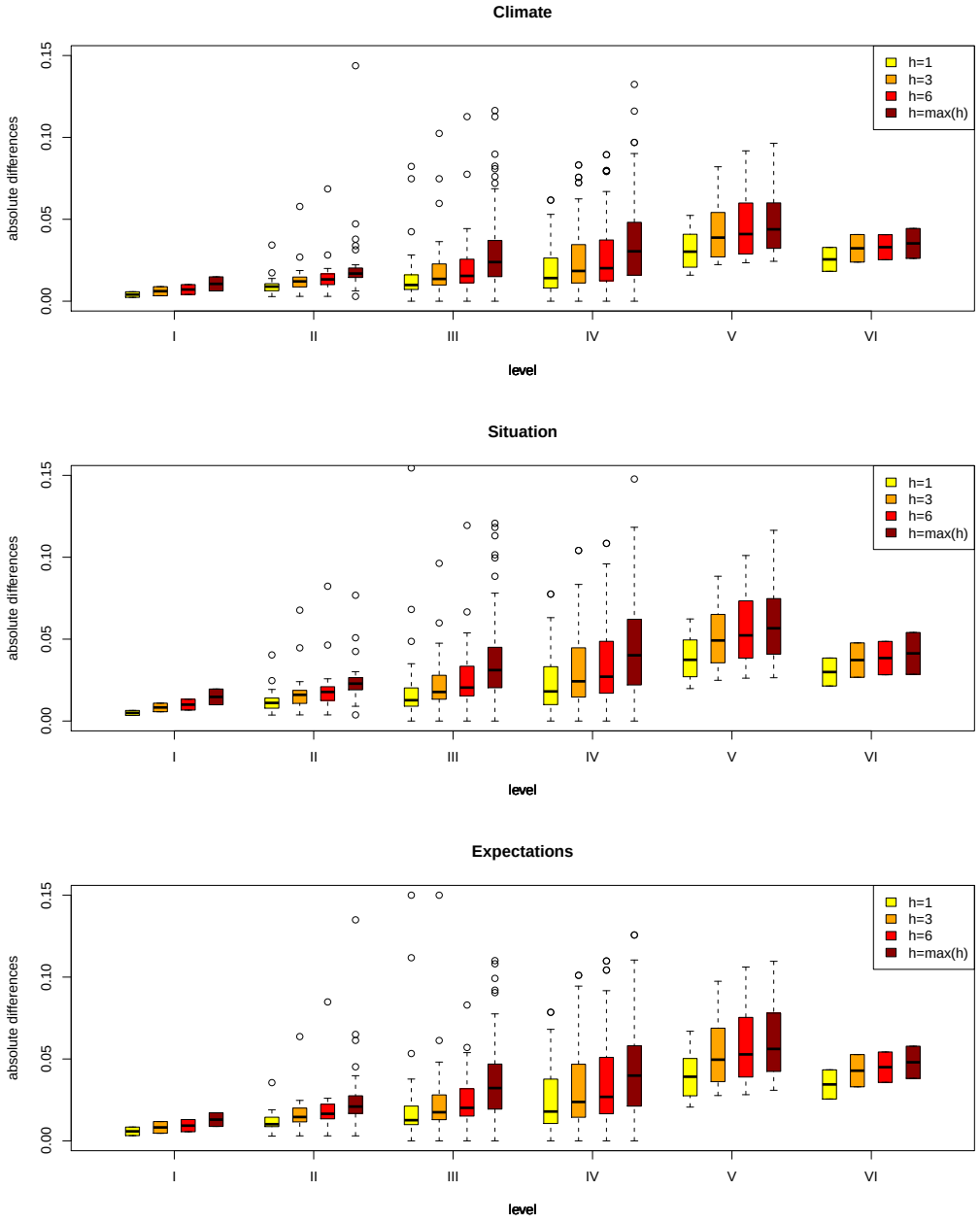

Figure 3: Boxplots for the distribution of the absolute differences between the original and the imputed indicators for different aggregation levels and horizons h.

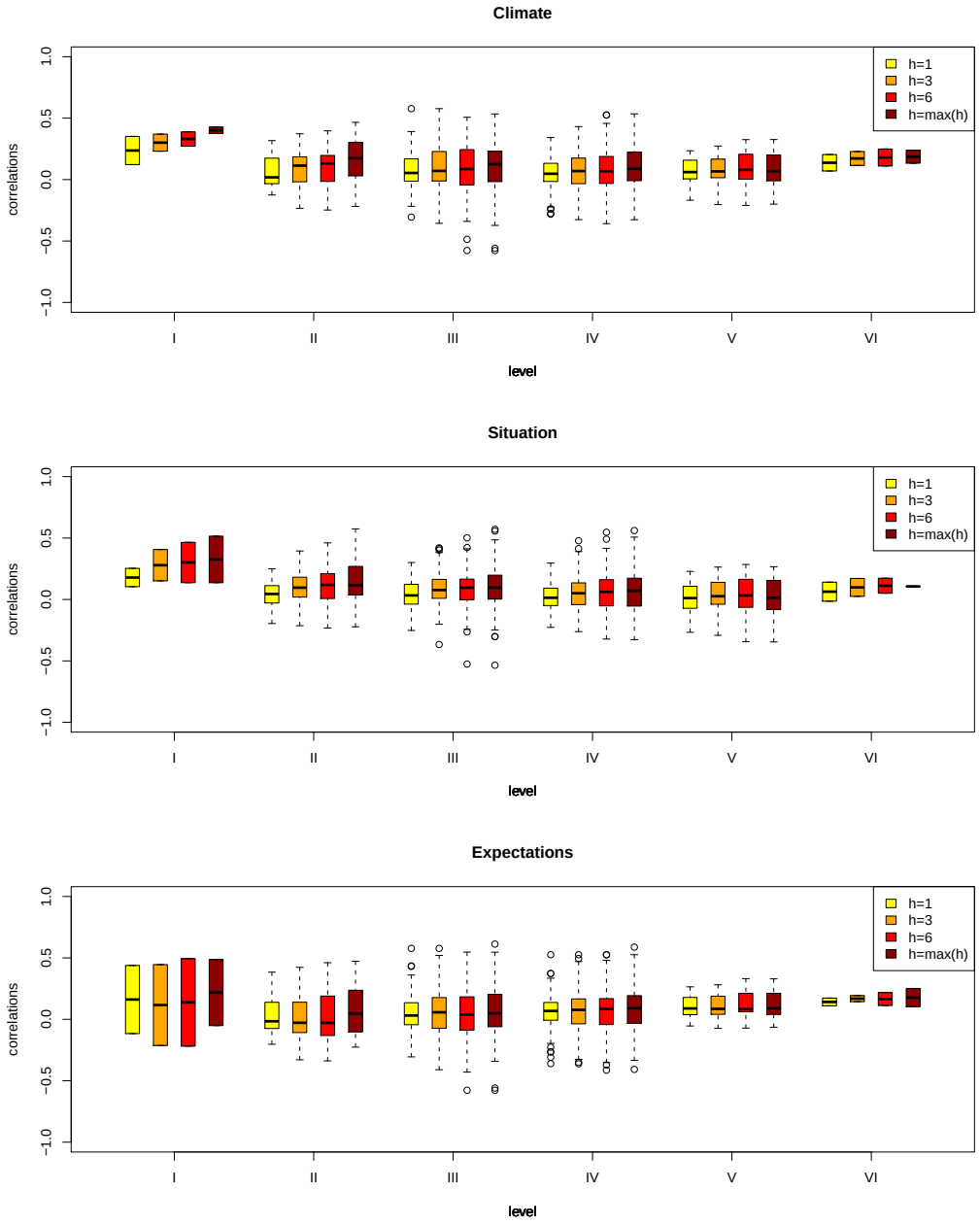

Figure 4: Boxplots for the distribution of ρ_{GDP} for different aggregation levels and horizons h.

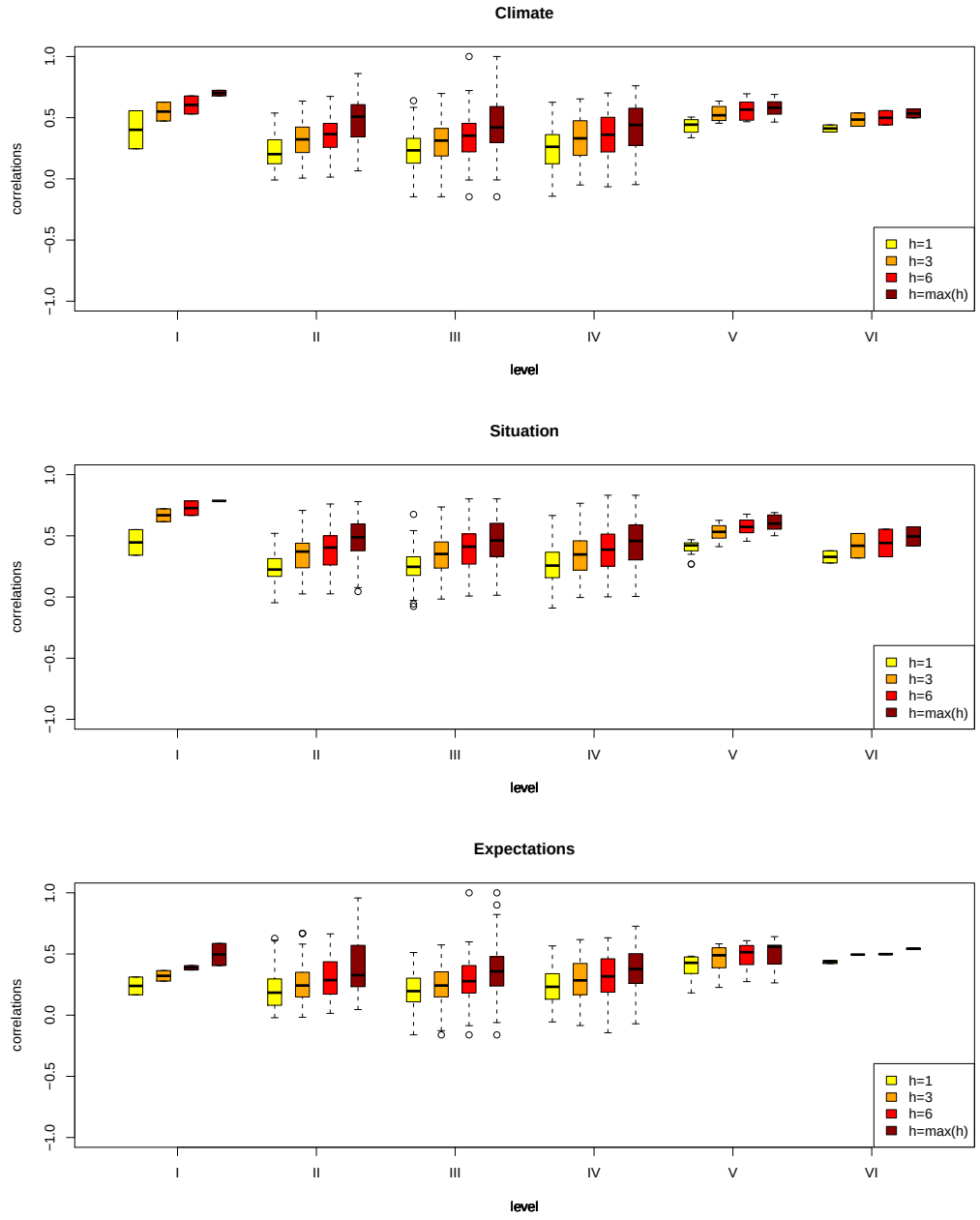

Figure 5: Boxplots for the distribution of ρ_{IND} for different aggregation levels and horizons h.

well as imputed) by

$$\tilde{y}_t = \frac{y_t - \bar{y}_t}{\sqrt{\text{Var}(y_t)}}$$

for any indicator y_t. Figure 9 shows the standardised indicators and their differences. It can easily be seen that the cycle dependence of the differences has vanished after standardisation. For example, for the standardised Ifo index and $h = 1$, $\rho^0_{GDP} = -0.114$ and $\rho^0_{IND} = -0.029$. Also for the subsectors we receive similar results, see Figures 7 and 8. Of course, it is highly discussible if a standardisation leads to a 'fairer' comparison between the original and the imputed indicators. However, a more distinctive comparison may be achieved by evaluating the difference in forecasting power which is done in the next Subsection.

4.2 Forecasting comparison

As noted in Section 1 we know that we can not perform a bias analysis without additional data. Although we showed in Section 3 that previous wave(s) contain enough information to produce relatively good estimates and the bias seems to be very small, we have to compare our indicators for their out-of-sample performance. Normally, the Ifo index is used as a leading indicator for the German GDP and the industrial production. Therefore, we will perform a 'horse race' between the original and the imputed indicators to test if forecasting performance of the imputed indicator is better than the original indicator. We consider a standard autoregressive distributed lag

(ADL) model

$$y_{t+h^*} = \alpha + \sum_{i^*=1}^{p} \phi_{i^*} y_{t+1-i^*} + \sum_{j^*=1}^{q} \theta_{j^*} x_{t+1-j^*} + \epsilon_t$$

with horizon $h^* = 1$. y_t denotes the quarter-on-quarter growth rate of the German GDP and x_t the aggregated quarterly Ifo index. For both variables, we allow the maximum of $p = q = 4$ lags and select the best model by AIC. We receive a RMSE ratio of 0.942, i.e. the imputed indicator leads, on average, to slightly better forecasts than the original indcator. To test whether this difference is statistical significant, we perform a Giacomini-White test (Giacomini and White,2006) which leads to a p-value of 0.348. Therefore, we can conclude that the imputed indicator does not lead to significant better forecasts.

5 Summary and discussion

In this paper, we developed different imputation strategies for a huge business survey with time-dependent latent process (the business cycle) and ordinal outcomes. Although the missing observations in our data set were caused in nearly all of the cases due to unit- and not item-nonresponse, we received good estimates by using the individual past as covariates for every unit. Also the predictive power of the imputations for runs of successive missings for a single firm was evaluated. But the analysis also showed that the strength of the imputation method is not always the same for every question in the survey. Questions regarded to the actual situation seem to be imputed with more certainty than questions with respect to future developments, which is as intuitive result as the latter inherent more uncertainty. After imputing missing observations with respect to different horizons of successive months of nonresponse, we recalculated the survey outcomes. The comparison with the original indicators showed that the bias is minimal, but generally increases with rising horizon and for indicators in sublevels. In addition, the selection bias seems to depend to a small extent on the business cycle, i.e. the latent variable. For the correlations with the cycle, we also found a similar effect as for the differences, so that the correlation rises when more values are imputed. To check our results with respect to forecasting power, we also performed a 'horse race' between the original and the imputed indicator. These results showed that the imputed indicator has a slightly better forecasting power, but this effect is not significant according to a Giacomini-White test.

However, our results do not hold if the indicators are standardised, in particular cycle dependence of the bias vanishes. This not concludes that a selection bias may not be present but confirms the results in Seiler and Wohlrabe (2012), who showed that the bias of a business cycle indicator is small even for very different patterns of NMAR and these only lead to a slight reduction in forecasting power. So, usage of such business cycle indicators (which base on surveys) for monitoring and forecasting the economy is secured under measurement error aspects due to nonresponse. Of course, the issue remains how such patterns may arise. Figure 6 shows the fractions of imputed values over time. For the business expectations, a small cycle dependence can be seen. Another very interesting issue is that LOCF seems to introduce a strong seasonal pattern but we notice that LOCF is no real model and therefore the results for this variable have to be interpreted with care. However, bad and equal states are imputed considerable more often than good states. The same, with exception of the seasonal pattern, regards to the business situation: For BS, we can see that the fractions of imputed values are very different according to the three states and are slightly cycle dependent across t which confirms the results of Harris-Kojetin and Tucker (1999) and Seiler (2010). This concludes that in general the nonresponse rate increases with the cycle but still firms are more likely respond if their situation and expectations are positive. But how does this fit to our results in Section 4? The pattern found here reduces the amplitudes of the indicators in boom times because more equal and negative values are imputed. But it also reduces the amplitudes in bad times as the imputed 'equal' values shift the indicator upwards. Figure 2 shows that the bias is, on average, positive,

especially for the business situation indicators. Therefore, it can be concluded that the selection bias, i.e. the differences between the three states, is more or less stable across time but the general decision to respond seems to be slightly correlated with the cycle. This bias leads to an overestimation of the indicators' amplitudes in extreme economic times (boom or recession). Since the level of these indicators is artifical and the forecasting performance is not reduced significantly, we conclude that the bias pattern found here is ignorable for this type of surveys and macro level results. However, micro level analyses as well as other surveys including quantitative information may be affected stronger by such biases.

Acknowledgements

The authors thank Kai Carstensen and Klaus Wohlrabe for their useful comments that have helped to improve this paper. The authors also thank Lisa Möst and Gunther Schauberger for their useful help.

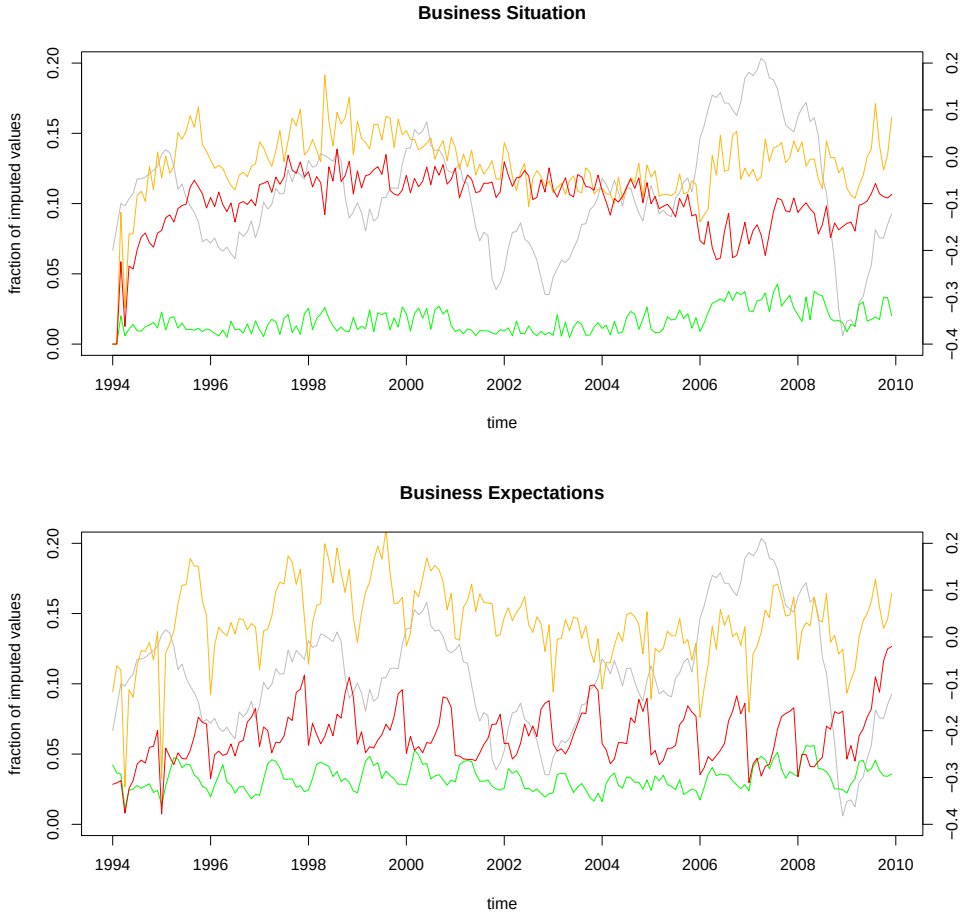

Figure 6: Fraction of imputed values (green line for '+', yellow line for '=' and red line for '-', left scale) in relation to the total number of contacted companies over time and the Ifo index (grey, right scale).

A Aggregation scheme of the Ifo index

As mentioned in Section 2, each company can give three possible replies to both, business situation and business expectations: a positive, a unchanged and an negative reply. The microdata aggregation has a tree structure according to the Classification of Economic Activities (edition 2008) from the German Federal Statistical of Office (Destatis, 2008). The lowest aggregation level evaluated in the Ifo Business Survey is IV in industry, VI in trade and V in construction (the Ifo index is level 0). We show the aggregation process based on the industry sector. For example, a company from manufacture of metal forming machinery (level IV) is part of manufacture of metal forming machinery and machine tools (level III) and part of manufacture of machinery and equipment n.e.c. (level II) and part of the industry sector (level I). There are 32 different groups on level II, 119 on level III, 171 on level IV, 24 on level V and 4 on level VI[10]. This process is equal for all variables measured on a 3-level Likert scale and all points in time.

We define $S_u^{IV}, u = 1, \ldots, U = 112$ as the u-th subsector on aggregation level IV. Each company can be clearly classified to one of these subsectors. We first count all weighted positive, unchanged and negative replies $\widetilde{m}^{+}_{t,S_u^{IV}}, \widetilde{m}^{=}_{t,S_u^{IV}}$ and $\widetilde{m}^{-}_{t,S_u^{IV}}$ in each subsector S_u^{IV}. The answers are weighted by the companies size, e.g. the answers of an industry firm with more than 500 employees get a weight of 15 whereas the answers of a company with less than 10 employees gets a weight of 1. Then, the number of weighted replies are scaled to the unit interval by dividing by $\widetilde{m}_{t,S_u^{IV}} = \widetilde{m}^{+}_{t,S_u^{IV}} + \widetilde{m}^{=}_{t,S_u^{IV}} + \widetilde{m}^{-}_{t,S_u^{IV}}$,

[10]Not all subgroups which occur in the German Classification of Economic Activities are calculated by the Ifo Institute, in particular no value for the whole trade sector is calculated.

so that

$$m^{\cdot}_{t,S_u^{IV}} = \frac{\widetilde{m}^{\cdot}_{t,S_u^{IV}}}{\widetilde{m}_{t,S_u^{IV}}}, \qquad m^{\cdot}_{t,S_u^{IV}} \in [0,1].$$

Then, the balance of u-th subsector S_u^{IV} is defined as

$$b_{t,S_u^{IV}} = (m^{+}_{t,S_u^{IV}} - m^{-}_{t,S_u^{IV}}), \qquad b_{t,S_u^{IV}} \in [-1,1], \tag{3}$$

i.e. to substract the fraction of negative from the fraction of positive replies. To calculate the balances for the next higher aggregation level, the fractions $m^{+}_{t,S_u^{IV}}, m^{=}_{t,S_u^{IV}}$ and $m^{-}_{t,S_u^{IV}}$ of replies are weighted, i.e.

$$m^{\cdot}_{t,S_v^{III}} = (m^{\cdot}_{t,S_1^{IV}}, \ldots, m^{\cdot}_{t,S_u^{IV}})' \omega_{S_v^{III}}$$

with $\omega_{S_v^{III}} = (\omega_{1,v}, \ldots, \omega_{U,v})', \omega_{u,v} \in [0,1], \sum_{u=1}^{U} \omega_{u,v} = 1$. Note that only $\omega_{u,v} > 0$ if $S_u^{IV} \in S_v^{III}$, i.e. if S_u^{IV} is subsector of S_v^{III}. The balances $b_{t,S_v^{III}}$ are just as calculated as in equation (3). The aggregation to level II, I and 0 is also carried out as described above. The index' value is obtained by scaling the balances to the average of the year 2005.

B Covariates for regression-based imputation

Industry	Construction	Trade
stock of inventories	construction activity vs. previous month	business volume vs. previous year
orders vs. previous months	construction activity in 3 months	feedstock (appraisal)
orders (appraisal)	constraints	prices vs. previous month
prices vs. previous months	constraints: lack of manpower	expected prices
expected production	constraints: lack of material	orders vs. previous year
expected domestic prices	constraints: weather conditions	
expected export trade	constraints: financing	
expected comercial operations	constraints: other reasons	
foreign orders (appraisal)	orders vs. previous month	
	orders (appraisal)	
	range of orders in months	
	prices vs. previous month	
	prime costs covering	
	expected prices	
	expected employees	
	industrial worker	
	employee	
	status of employee's illness in %	

Table 4: Covariates included in $\boldsymbol{x}_{t-1}^{sec}$ for the different sector models

C Results for standardised indicators

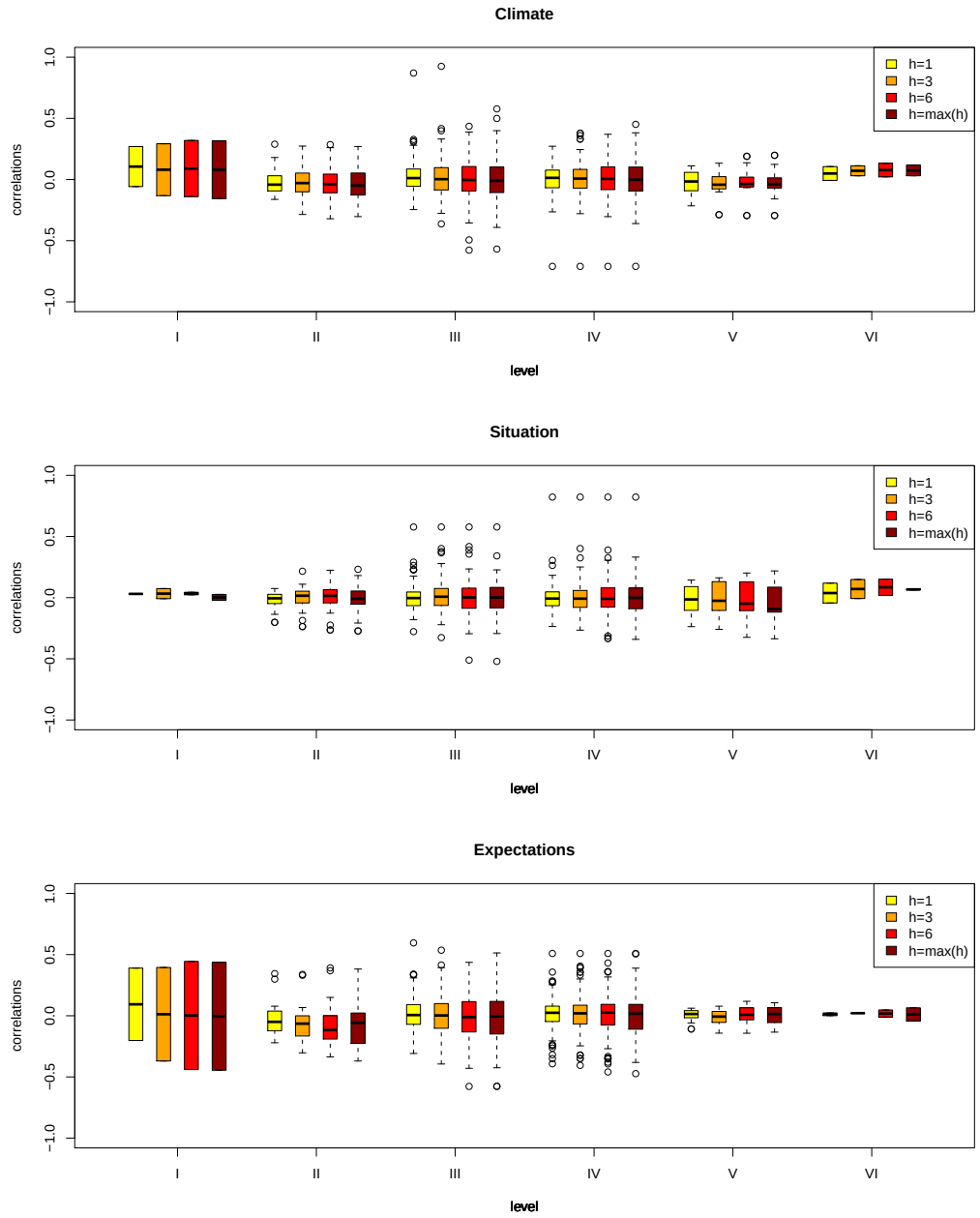

Figure 7: Boxplots for the distribution of ρ_{GDP} for the *standardised* original and imputed indicators, different aggregation levels and horizons h.

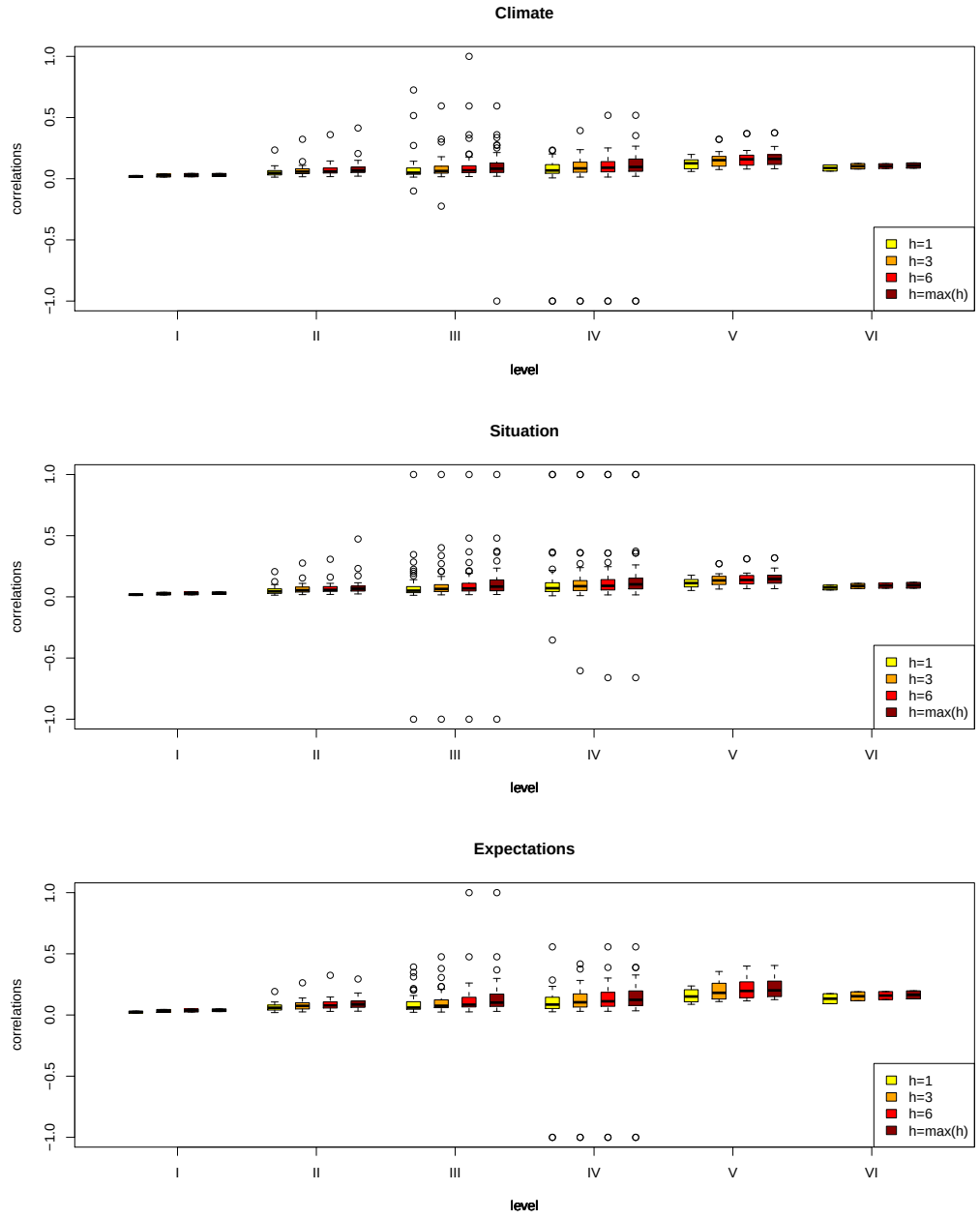

Figure 8: Boxplots for the distribution of ρ_{IND} for the *standardised* original and imputed indicators, different aggregation levels and horizons h.

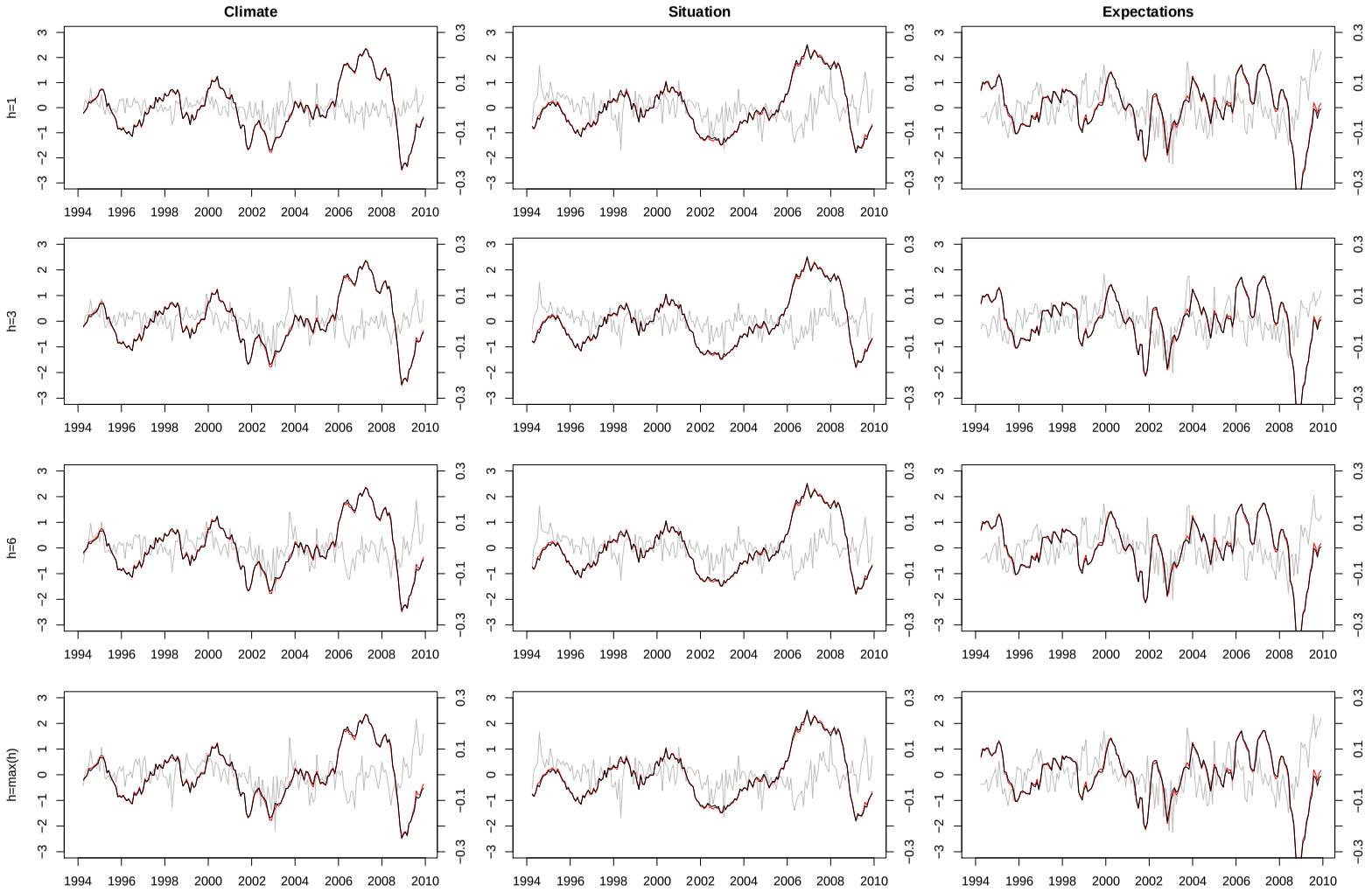

Figure 9: Original (black) and imputed (red) *standardised* Ifo indicators and their difference (gray, right scale)

References

Abberger, K. and Wohlrabe, K. (2006). Einige Prognoseeigenschaften des ifo Geschäftsklimas - Ein Überblick über die neuere wissenschaftliche Literatur. *ifo Schnelldienst*, 59(22):19–26.

Anderson, O. (1951). Konjunkturtest und Statistik. *Allgemeines Statistisches Archiv*, 35:209–220.

Anderson, O. (1952). The business test of the IFO-Institute for Economic Research. *Revue del'Institute International de Statistique*, 20:1–17.

Becker, S. O. and Wohlrabe, K. (2008). Micro Data at the Ifo Institute for Economic Research - The "Ifo Business Survey", Usage and Access. *Journal of Applied Social Science Studies*, 128(2):307–319.

Cameron, A. C. and Trivedi, P. K. (2005). *Microeconometrics. Methods and Applications.* Cambridge University Press.

Chen, J. and Shao, J. (2000). Nearest Neighbor Imputation for Survey Data. *Journal of Official Statistics*, 16(2):113—-131.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46.

Cook, R. J., Zeng, L., and Yi, G. Y. (2004). Marginal Analysis of Incomplete Longitudinal Binary Data: A Cautionary Note on LOCF Imputation. *Biometrics*, 60(3):820–828.

Destatis (2008). *Classification of Economic Activities, Edition 2008*. Federal Statistical Office of Germany.

Engels, J. M. and Diehr, P. (2003). Imputation of missing longitudinal data: a comprison of methods. *Journal of Clinical Epidemiology*, 56:968–976.

European Union (2006). Joint Harmonised EU Programme of Business and Consumer Surveys. *Official Journal of the European Union*, 49(C 245):5–8.

Finch, W. H. (2010). Imputation Methods for Missing Categorical Questionnaire Data: A Comparison of Approaches. *Journal of Data Science*, 8:361–378.

Giacomini, R. and White, H. (2006). Tests of Conditional Predictive Abilitys. *Econometrica*, 74(6):1545–1578.

Goldrian, G., editor (2007). *Handbook of survey- based business cycle analysis*. Edward Elgar Publishing.

Graham, J. W., Olchowski, A. E., and Gilreath, T. D. (2007). How Many Imputations are Really Needed? Some Practical Clarifications of Multiple Imputation Theory. *Preventative Science*, 8:208–213.

Harris-Kojetin, B. and Tucker, C. (1999). Exploring the Relation of Economical and Political Conditions with Refusal Rates to a Government Survey. *Journal of Official Statistics*, 15(2):167–184.

Honaker, J. and King, G. (2010). What to do About Missing Values in Time Series Cross-Section Data. *American Journal of Political Science*, 54(2):561–581.

Honaker, J., King, G., and Blackwell, M. (2011). *Amelia II: A Program for Missing Data*.

Janik, F. and Kohaut, S. (2011). Why don't they answer? - Unit non-response in the IAB Establishment Panel. *Quality & Quantity*. to appear.

King, G., Honaker, J., Joseph, A., and Scheve, K. (2001). Analyzing incomplete political science data: An alternative algorithm for multiple imputation. *American Political Science Review*, 95(1):49–69.

Kleinke, K., Stemmler, M., Reinecke, J., and Lösel, F. (2011). Efficient ways to impute incomplete panel data. *AStA Advances in Statistical Analysis*. to appear.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Little, R. J. A. and Rubin, D. (2002). *Statistical Analysis with Missing Data*. Wiley.

Manski, C. (2003). *Partial Identification of Probability Distributions*. Springer.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society, Series B*, 42:109—142.

Nardo, M. (2003). The quantification of qualitative survey data: A critical assessment. *Journal of Economic Surveys*, 17(5):645–668.

OECD (2003). *Business Tendency Surveys - A Handbook*.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley.

Saha, C. and Jones, M. P. (2009). Bias in the last observation carried forward method under informative dropout. *Journal of Statistical Planning and Inference*, 139(2):246–255.

Schafer, J. (1997). *Analysis of Incomplete Multivariate Data*. Chapman and Hall.

Schunk, D. (2008). A Markov chain Monte Carlo algorithm for multiple imputation in large surveys. *AStA Advances in Statistical Analysis*, 92(1):101–114.

Seiler, C. (2010). Dynamic Modelling of Nonresponse in Business Surveys. Ifo Working Paper 93, Ifo Institute.

Seiler, C. and Wohlrabe, K. (2012). Surveys, Nonresponse, and the Business Cycle. CESifo Working Paper, Ifo Institute. to appear.

Theil, H. (1952). On the shape of economic microvariables and the Munich business test. *Revue del'Institute International de Statistique*, 20:105–120.

Woolley, S. B., Cardoni, A. A., and Goethe, J. W. (2009). Last-observation-carried-forward imputation method in clinical efficacy trials: review of 352 antidepressant studies. *Pharmacotherapy*, 29(12):1408–1416.