SPECIALISTS IN
EMPIRICAL ECONOMIC
RESEARCH

GWS DISCUSSION PAPER 2020/1

Back to the future.

Counterfactual Scenarios – Challenges, Methodology and an Empirical Test

Anke Mönnig

Frank Hohmann

Impressum

AUTHORS

Anke Mönnig

Tel: +49 (541) 40933-210, E-Mail: moennig@gws-os.com

Frank Hohmann

Tel: +49 (541) 40933-130, E-Mail: hohmann@gws-os.com

TITLE

Back to the future. Counterfactual Scenarios – Challenges, Methodological and an Empirical Test

PUBLICATION DATE

DISCLAIMER

FUNDING

The results at hand have been prepared in the course of a research project for the German Federal Ministry of Research and Education (BMBF FKZ: 01DN16030).

PUBLISHER OF THE GWS DISCUSSION PAPER SERIES

Gesellschaft für Wirtschaftliche Strukturforschung (GWS) mbH

Heinrichstr. 30

49080 Osnabrück (Germany)

ISSN 1867-7290

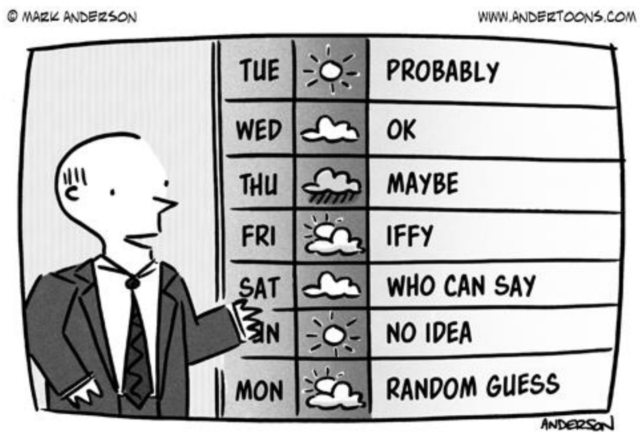
TABLE OF CONTENTS

1 INTRODUCTION

Looking in the future is a challenge. Most if not all may even say it is impossible. However, forecasting has become an important part for policy planning. Economic forecasts offer guidance under conditions. They are used and /or produced by politicians, researchers, companies, associations or unions. They enter decision making processes and have an impact on state budget, consumption decision, personnel strategies – just to name a few.

Ex-ante (policy) impact assessment (IA) is a forward-looking concept that has to deal with a lot of unknowns (e.g. natural disasters). The predicted impact is only valid within a certain framework or set of assumptions. However, it allows to pass judgements on the effectiveness and efficiency of planned measures. In many countries, impact assessments "has become an important tool for assisting policy makers in their decision-making process" (Großmann et al. 2016: 13).

Due to its importance of (ex-ante) impact assessment, the questions arises regularly whether the forecasted results are robust. Or put differently: How good is the forecast? One method to answer this is to apply counterfactual forecasts.



"And now the 7-day forecast..."

Such counterfactual scenarios or ex-post scenarios are, however, challenging to model. There can be two reasons for going "back to the future" and for facing this challenge: First, to test the accuracy of a forecasting model, or, second, to study the efficiency of an already implemented policy. While the first reasons produces a "first order" ex-post simulation, the second reason is a "second order" ex-post simulation where first order results are used as a baseline.

Only first-order simulations can be compared to the already known real world. With some diagnostic checks – such as mean, relative or squared error tests – the model forecasting performance can be evaluated. However, the question is not if there are error terms to be observed but how big they are. Second-order scenarios can only be compared to first-order scenarios, not to actual data.

In this paper, we introduce how to perform ex-post forecast with a macroeconometric input-output model. We take the example of COFORCE which has been developed to forecast the Chilean economy until 2035 (Mönnig & Bieritz 2019).

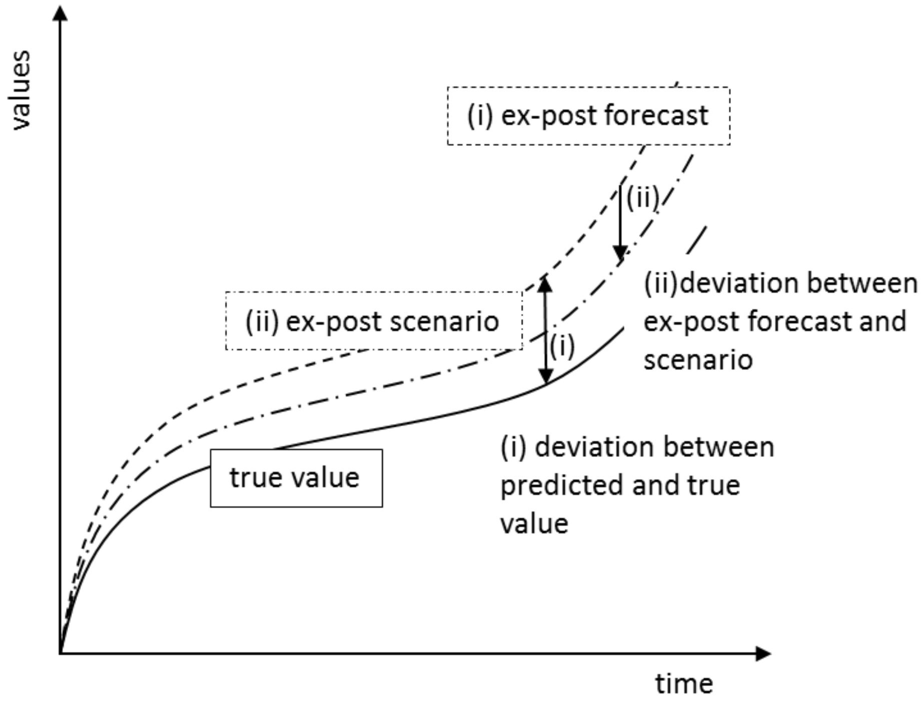
The remainder of the paper is structured as follows: First, an overview about the challenges concerned with ex-post simulations is given. Then, the methodological approach is described. Next, an ex-post simulation is performed on the model COFORCE. The paper concludes with the main findings.

2 THE CHALLENGE

Counterfactual scenarios are powerful and helpful tools for doing two things: (i) to evaluate forecasts (ex-post forecast) and (ii) to evaluate past policy measures (ex-post scenario). Figure 1 summarizes the differences between true (historical given) values, ex-post forecasts (refers to (i)) and ex-post scenarios (refers to (ii)).

The second intended use – evaluation of past policy measures – is more often applied. In such a context, counterfactual scenarios are also called ex-post (policy) impact assessments (ex-post scenario). They are used for evaluating already implemented (policy) measures. In Figure 1 this is referred to as ex-post scenario. The challenge is to find a credible approximation to what would have happened in the absence of the policy measure, and to compare it with what has actually occurred (European Commission 2013). The difference is the estimated impact of the intervention, on the particular outcome of interest (e.g. GDP). This method is applied for instance for questions related to energy policy measures and the question how the economy would look alike if these measures hadn't been implemented. For example, if the energy price privileges had been abolished (FISI & ECOFYS 2015) or if the extension of renewable energy hasn't been taken place (Lutz et al. 2014).

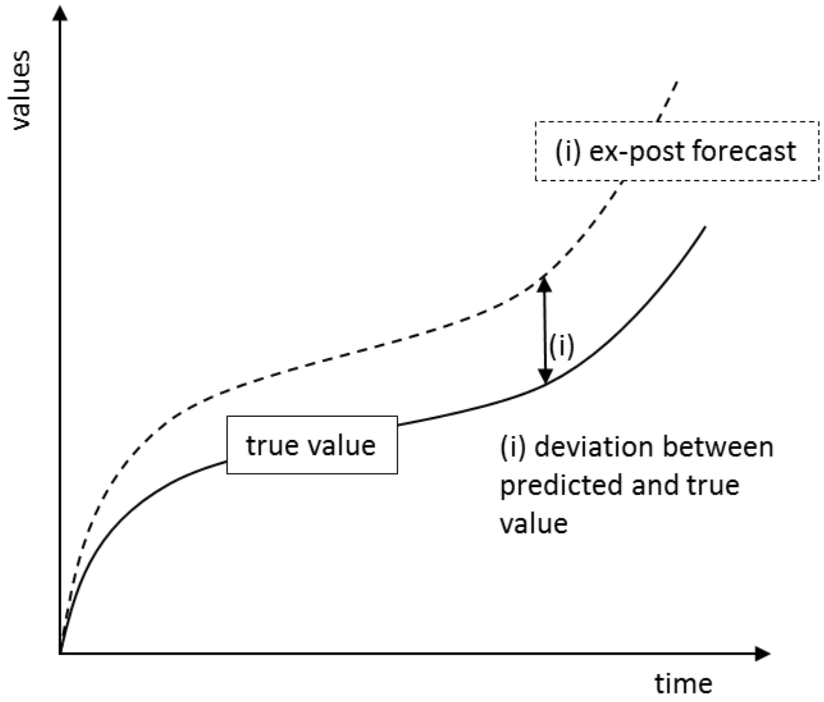
Figure 1: Difference between true values, ex-post scenario and ex-post forecast



Source: own drawing

More seldom, counterfactual scenarios are used for evaluating the robustness of the forecasting model (compare Figure 2). Ash et al. (1990) have evaluated the OECD forecasts for the period 1968-1987. Baumgartner (2002) has compared the forecasting robustness of two projection models in Austria for the period 1978-1999. The reason why this is not been done more regularly may be because the likelihood of failing to forecast true values is high. Forecasting models are vulnerable to lagging behind real time.

Figure 2: Difference between true value and ex-post forecast



Source: own drawing

Trying to answer how good a forecast is demands an intellectual twist: in order to evaluate a forecast the true values of the forecast must be known. This time warp can be resolved by applying for instance counterfactual forecasts. Counterfactual forecasts pretend to go back to the past and predict the future again. Simply put: if the forecasted values are equal to the observed values, the forecasting method was good. If they don't match, the forecasting method was bad.

However, interpreting counterfactual forecasts is more complicated than this. When looking more closely at the subject, the following mind traps emerge:

1. Unlike weather forecasts – which have no impact on actual weather conditions – economic forecasts can have an impact on the behavior of economic actors. Only if a forecast has no real impact, the comparison of predicted and true values are a good proxy for judging a forecast. However, there is something that economists like to call self-fulfilling-prophecy or refer to as Lucas critique. This is a case when a forecast has an influence – a negative or a positive – on the real economy. In such a case, a comparison of a forecast with true values is not a good indicator for judging a forecast.

2. Economic forecasts are forecasts under conditions. A set of assumptions are necessary to perform a forecast. In most computational models – independent of what type –, such assumptions are for instance interest rates, world trade dynamics, oil prices, population development, exchange rates. These assumptions are exogenously set, but have strong influence on the results of the economic forecast without being affected by the forecast itself. A wrong set of these assumptions – for example a different population development due to a sudden strong increase in

migration – leads to a mismatch between forecasted and true values.

3. Keynes' hypothesis was that "theory must be confirmed if the data and statistical methods are employed correctly" (Woschnagg & Cipan 2002). His claim was that true values have to be met by a forecasting model, if the forecasting model is correct. Otherwise, there is something wrong with the model. Therefore, the economic philosophy that frames the computational model determines and defines the future output. It refers to the economic explanatory context in which the forecast is embedded. A forecast may look good – compared to the true values – but only by "accident". By contrast, if the true values are not met with the a correct model, some explanatory elements are missing.

4. The judgement about the robustness of a forecast may also vary between different interested parties (Baumgartner 2002: 702). Baumgartner (2002) differs between a symmetric and asymmetric loss function. A symmetric loss function exists when the positive and negative deviation to the true value are judged as evenly wrong. Policy makers, however, may have an asymmetric loss function, because they evaluate the positive deviation higher than a negative deviation to the true values: a non-predicted recession may cause more harm than an unforeseen economic upswing.

3 THE METHODOLOGY

3.1 TECHNICAL SETTING

Preparing a consistent dataset for an econometric model is already a challenge which becomes even more demanding if the model should also be used for performing counterfactual scenario analysis.

One key feature of an elaborated econometric model is the use of up-to-date time series which may vary with respect to their individual time span. In most cases a model builder has to deal with the problem that the dataset consists of time series with different start end dates. The most common cause of this problem stems from the fact that a dataset is typically compiled from different data sources with individual dataset characteristics such as update cycle. Even data from one provider may have different update cycles and/or time spans for various reasons.

The technical setup of an econometric model has to consider this problem adequately. First, measures are needed to keep historic data from being overwritten unintentionally. Second, updating a model with newly available data should be a straightforward procedure. Performing counterfactual scenario analysis increases the need for effective measures as it introduces another variant of a model by shifting the calculations into the past.

The most straightforward option to keep data from being overwritten is to use a conditional statement which is part of almost every programming language or model building environment. Assuming that the time series for "Changes in inventories" *CIES* is available up to 2013, its calculation could be guarded by the following conditional statement in order to prevent unintended data corruption:

```
if(t > 2013)
     CIES[t] = …  //some calculation
```

where *t* denotes the current date of calculation at model runtime.

Although this approach is easy to implement, it proves to be error-prone if a dataset gets an update: The full model code needs to be carefully revised to reflect the new start date for each of the updated variables. The situation is even worse with respect to the calculation of counterfactual scenarios: The model code must be duplicated to allow for calculation of parts of the historical time span, e.g.

```
if(CIES[t] > 2010) // start date of counterfactual scenario
     CIES[t] = …  //some calculation
```

Obviously, this approach introduces a serious problem. Every change to one model type (e.g. forecasting model) must be duplicated in the other model type (e.g. counterfactual analysis model) to ensure consistency across both models. If, for example, the counterfactual analysis reveals that the behavioral equation for a certain variable should be replaced, this altered equation needs to be copied over to the forecasting model to keep both model versions in sync.

A much better approach is to use placeholders for the start dates in the model code which the are populated by the model engine at runtime. The best implementation of this approach depends on the programming language or model building environment in use.

If the model calculation engine does not provide individual time series properties like start date or if the engine does not allow for adding user-defined attributes to model variables, the problem can be solved by implementing a generator program which creates a list of necessary placeholders with respect to the model type on the fly. This list can then be injected into the model, e.g.

```
const CIES_lastData = 2013

const GDPD_lastData = …
…
```

In the model code, the calculation of CIES now reads as

```
if(t > CIES_lastData)
     CIES[t] = …  //some calculation
```

With this version, the model code is not affected by changes to the time span of the calculation and/or individual start dates anymore and thus makes switching between the different model types almost straightforward.

If the programming language or model building environment supports the processing of additional input files, switching between the different model types becomes possible without the need for updating any model source code file. This is especially useful for languages which require recompilation of changes to the source codes. The *lastData* values will be stored in external data files along with the variable names. The values are then retrieved and assigned at model startup by a dedicated initialization routine.

3.2 MEASURES FOR DIAGNOSIS

The most straightforward way to evaluate the robustness of a forecast is comparing the estimated with the true values. However, this can be done in many ways: comparing absolute or relative deviations, squared deviations, etc.. Accordingly, the judgement about the accuracy of the forecast differs. A deviation can be small in absolute terms, but large in relative terms, and vice versa. Not all error measures are sensible to use for the same thing. It is also important to realize that the choice of diagnostic check determines the degree to which a forecast can be interpreted as robust.

It is important to know what measures exist to evaluate forecasts, when they are useful to use and how they are interpreted. A very good overview is given for example in Andres & Spiwoks (2000).

Inspired by Andres & Spiwoks (2000), a selected number of measures is introduced in the following that are assigned to the following headings: (1) simple error measures, (2) cumulative error measures, (3) relative error measures, (4) squared error measures and (5) comparison with naïve forecasts.

The following applies to all introduced error measures: The smaller the calculated statistics, the better the forecast.

An error measure e is interpreted as the difference between predicted $\hat{x}$and true value x.

$$e = \hat{x} - x$$

(1) Simple error measures

Such measures are very simple in its structure and are easy to interpret. Although less complex, they can give a first good sign to the accuracy of a forecast:

LFE	Largest forecast error	$max(\hat{x} - x)$	Shows only the largest deviation
SFE	Smallest forecast error	$min(\hat{x} - x)$	Shows solely the smallest deviation
ME	Mean error	$\frac{1}{T}\sum_{t=1}^{T}(\hat{x} - x)$	Shows the average deviation to the true value. The closer to zero, the smaller the global over- or underestimation of the forecast error A positive sign indicates that the forecast (tends to) underestimate the actual development.
MPE	Mean positive error	$\frac{1}{T}\sum_{\substack{t=1 \\ (\hat{x}-x)\geq 0}}^{T}(\hat{x} - x)$	Shows the average overestimation of the forecast error
MNE	Mean negative error	$\frac{1}{T}\sum_{\substack{t=1 \\ (\hat{x}-x)\leq 0}}^{T}(\hat{x} - x)$	Shows the average underestimation of the forecast error

MAE	Mean absolute error	$\frac{1}{T}\sum_{t=1}^{T}\lvert(\hat{x}-x)\rvert$	Shows the average distance to the true value. Over- and underestimation are not balanced. It measures the accuracy of a forecast. Errors are weighted linearly.

$T = total\ observation, t = single\ observation, x = true\ value, \hat{x} = predicted\ value$

(2) Cumulative error measures

Cumulative forecast error can be used if it is interesting to know to what extent forecasts tend to over- or underestimate.

COE	Cumulative overestimation error	$\sum_{\substack{t=1 \\ (\hat{x}-x)\geq 0}}^{T}(\hat{x}-x)$	Shows the total cumulative overestimation over a given period
CUE	Cumulative underestimation error	$\sum_{\substack{t=1 \\ (\hat{x}-x)\leq 0}}^{T}(\hat{x}-x)$	Shows the total cumulative underestimation over a given period
CAE	Cumulative absolute error	$\sum_{t=1}^{T}\lvert(\hat{x}-x)\rvert$	Shows the total cumulative error over a given period

$T = total\ observation, t = single\ observation, x = true\ value, \hat{x} = predicted\ value$

(3) Relative error measures

Relative forecasting error measures are used if a standardization in the comparison of variables is necessary. Relative measures are especially useful and superior to absolute error measures (compare (1)) especially if longer time series are observed because they take into account the underlying (economic) situation. In times of economic stagnation – for instance – an absolute error term of two percent is quite large, whereas it is of less relevance in a period of strong economic growth of maybe two percent.

MRE	Mean relative error	$\sum_{t=1}^{T}\frac{(\hat{x}-x)}{x}$	Shows the average deviation to the true value relative to the true value. The closer to zero, the smaller the global over- or underestimation of the forecast error. Multiplied by 100 gives the average percentage deviation to the true value
MRAE	Mean relative absolute error	$\sum_{t=1}^{T}\left\lvert\frac{(\hat{x}-x)}{x}\right\rvert$	Shows the relative average distance to the true value. Over- and underestimation are not balanced. Multiplied by 100 gives the average percentage distance to the true value

$T = total\ observation, t = single\ observation, x = true\ value, \hat{x} = predicted\ value$

(4) Squared error measures

The deviation between predicted and true values is squared which means that the larger the deviation, the larger the error term.

MSE	Mean squared error	$\frac{1}{T}\sum_{t=1}^{T}(\hat{x}-x)^2$	The deviations of forecasted and true values are squared, summed and divided by the number of observations. This prevents positive and negative deviations from being balanced. There is a disproportionate weighting of large and small deviations. It measures the accuracy of a forecast. Errors are weighted quadratically.
MSRE	Mean squared relative error	$\frac{1}{T}\sum_{t=1}^{T}\left(\frac{(\hat{x}-x)}{x}\right)^2$	Is the relative equivalent to the mean squared error.
RMSE	Root mean squared error	$\sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(\frac{(\hat{x}-x)}{x}\right)^2}$	Especially used in anglo-saxon literature.

$T = total\ observation, t = single\ observation, x = true\ value, \hat{x} = predicted\ value$

The mean squared error (MSE) can be further decomposed in inequality shares: distortion (UM) and regression component (UR) and the remaining distribution component (UD). All three shares must add up to one. The distribution component (UD) should be close to one, the other two components close to zero.

$$UM + UR + UD = 1$$

(5) Comparison with naïve forecasts (Theil's inequality coefficient)

Comparing projected and true values is a necessary but not a sufficient method in evaluating a forecast accuracy because absolute or squared error measures do not include any information about how difficult it is to predict a particular variable (Baumgartner 2002: 705). Variables with less variation are easier to predict than those with greater variation.

Therefore, a further evaluation measure is the comparison of projected value with a naive forecast and compare both estimated values with the true value.

MRWnF	Mean relative error weighted with a naïve forecast	$\frac{\sum_{t=1}^{T}(\hat{x}-x)}{\sum_{t=1}^{T}(\bar{x}-x)}$	The naïve forecast implies the lowest level of forecasting quality. If the forecast is not better than the naïve forecast, it means that the projected relations don't

			make sense. If the measure equals 1, the average over and underestimation of the forecast is equal to the one of the naïve forecast.
MRAWnF	Mean relative absolute error weighted with a naïve fore-cast	$\frac{\sum_{t=1}^{T} \lvert(\hat{x}-x)\rvert}{\sum_{t=1}^{T} \lvert(\bar{x}-x)\rvert}$	It is the absolute equivalent of the mean relative error weighted with a naïve fore-cast. The absolute approach avoids the bal-ancing of over and underestimations. The absolute distance to the true value is leading. If the measure equals 1, the forecast is as equally good or bad as the naïve fore-cast. Is the measure >1, the forecast is worse than the naïve forecast Is the measure <1, the forecast is better than the naïve forecast.

$T = total\ observation, t = single\ observation, x = true\ value, \hat{x} = predicted\ value$

The introduced measures in the table above can be calculated with all other types of error measures. Often, root-mean-squared-error (RMSE) is used for comparing two forecasts (Baumgartner 2002: 705). A good forecast shows values for Theil's inequality coefficient that are significantly smaller than 1.

4 THE EMPIRICAL TEST

The research project “Development of sustainable strategies in the Chilean mining sector through a regionalized national model” (BMBF FKZ: 01DN16030) inquires the socio-economic impacts of copper on the Chilean economy. The project is government-funded by the BMBF[1] and supports the cooperation and exchange of knowledge between the Chilean team of Prof. Dr. Aroca and the GWS, a German company for empirical economic research. In September 2016, the project started.

Part of the project was to develop three scenarios to test the vulnerability of Chile’s economy to copper. The three scenarios addressed different aspects of the economy: the first scenario “energy” focused on the effect of a changing energy mix on copper production. The second scenario addressed the high relevance of copper export to the Chilean economy and the effects of an world demand shift away from Chile and towards Peru as major copper supplier. The third scenario is a counterfactual scenario that aimed to quantify the

[1] German Federal Ministry of Education and Research

past effect of regime shift in the taxing system. Whereas the first two scenarios are scenarios that address future changes (ex-ante perspective), the last scenario aimed to separated past effects (ex-post perspective).

The first order ex-post forecast is used as the baseline for the ex-post tax scenario. On the example of COFORCE – a macroeconometric model developed to forecast the Chilean economy until 2035 (Mönnig & Bieritz 2019) – the ex-post forecast is evaluated by using some – not all – measures introduced in section 3.2.

4.1 CALIBRATING THE MODEL

The ex-post forecast is calculated for the years 2009 to 2013. The total observation T equals 5 years.

All exogenous variables remain exogenous. These are population, interest rate, exchange rate, nominal exports and import prices.

All endogenous variables remain endogenous, except for a few.

- Consumption of non-profit organizations. This is a very small area only.
- Government expenditures (nominal). Government expenditures depend on policies that cannot be foreseen by the model. Past policies are difficult to capture empirically. Policies (e.g. changes in taxation or regulation, fiscal packages) that happened in the past are usually insufficiently mirrored in the coefficients.
- Nominal inventories. No empirically explanation is possible.
- Two producer prices. The model is very sensitive to changes in producer prices for *extractive fishing industry* and *other manufacturing industries*.

4.2 DIAGNOSTIC CHECKS

The diagnostic checks can be run on all variables in the model. In this paper we concentrate only on the aggregates of real GDP and its components as well as on nominal production on sectoral level.

The error measures are based on the difference between growth rates – not on the difference between absolute values. The total number of observation is $T = 5$.

Table 1 shows a collection of selected simple and relative error measures. The smaller the values, the better the forecast.

Mean error (ME) shows the general tendency of the forecast. A negative GDP value indicates that real GDP is – in tendency – underestimated. In average, the true value is higher than the predicted value. The same observation holds for real private consumption, real investment and imports. Export and consumption of NGOs are generally underestimated. However, both values are relatively close to zero, which indicates, that consumption of NGOs and real exports show a small tendency of over- or underestimation

The average distance of the forecasted and true values is given with MAE. The average distance of 2.44 for real GDP is relatively high. Only real investment has a MAE of below

2 which means that the average distance to the true value is less strong.

Looking at consumption of NGOs, the different between ME and MAE becomes evident: While a ME of 0.55 is close to zero which indicates a good average proxy to the true value, the high MAE value of 3.45 indicates that consumption of NGOs is largely over- and underestimated over time. MAE does not balance positive and negative deviations. In contrary: real investment holds the same value for ME and MAE which indicates that between 2009 and 2013, real investments are constantly underestimated.

Looking at MRE, all values – despite real exports – are close to zero. This is a good indicator for a good forecast. Multiplied by 100, MRE is the average percentage difference to the true values. The percentage difference for real GDP is -30%. On average, the forecast only meets the true values by 70%. The percentage difference is especially high for real exports and especially small for real consumption of NGOs and for real investments.

In total, the error measures of Table 1 suggest that the forecast is generally underestimated with relatively high distances to the true value. Among GDP components, real investments seem to be pretty robust in their error measures. In all three applied error measures, it shows the best and most stable results.

Table 1: Selected simple and relative error measures

	Mean error (ME)	Mean absolut error (MAE)	Mean relative error (MRE)
	$\frac{1}{T}\sum_{t=1}^{T}(\hat{x}-x)$	$\frac{1}{T}\sum_{t=1}^{T}\lvert(\hat{x}-x)\rvert$	$\sum_{t=1}^{T}\frac{(\hat{x}-x)}{x}$
	Mean deviation of growth rates 2009-2013	Absolute mean deviation of growth rates 2009-2013	Mean relative deviation of growth rates 2009-2013
	The smaller the better the forecast		
GDPTR	-1,52	2,44	-0,29
HCESR	-2,04	3,45	-0,26
CPONR	0,55	3,13	0,16
GCESR	0,00	0,00	0,00
GICNR	-1,29	1,29	-0,23
EGGSR	0,18	3,12	-3,15
IGSSR	-1,68	2,08	-0,19

Legend: GDPTR – real gross domestic product; HCESR – real private consumption; CPONR – real consumption of non-profit organisations; GCESR – real government consumption; GICNR – real gross investment; EGGSR – real exports; IGSSR – real imports

Table 2 turns to squared error measures. High deviations are weighted stronger than small deviations. As the values of MSE are in general much higher than the ones observed in Table 1, this indicates that high deviations can be observed more regularly. This is especially the case for real private consumption and also for real consumption of NGOs. Also real exports and imports have a much larger MSE than ME or MAE. Again, real investments are an exception. With an MSE of 1.97, the squared error measure of real investments is not significantly higher than in Table 1. This indicates that deviations to the true value are not very high.

The decomposition of MSE confirms the not very satisfactory results of MSE. In general, the proportion forecast error (UM and regression error (UR) are small, but not small enough to allow for a robust decomposition of MSE.

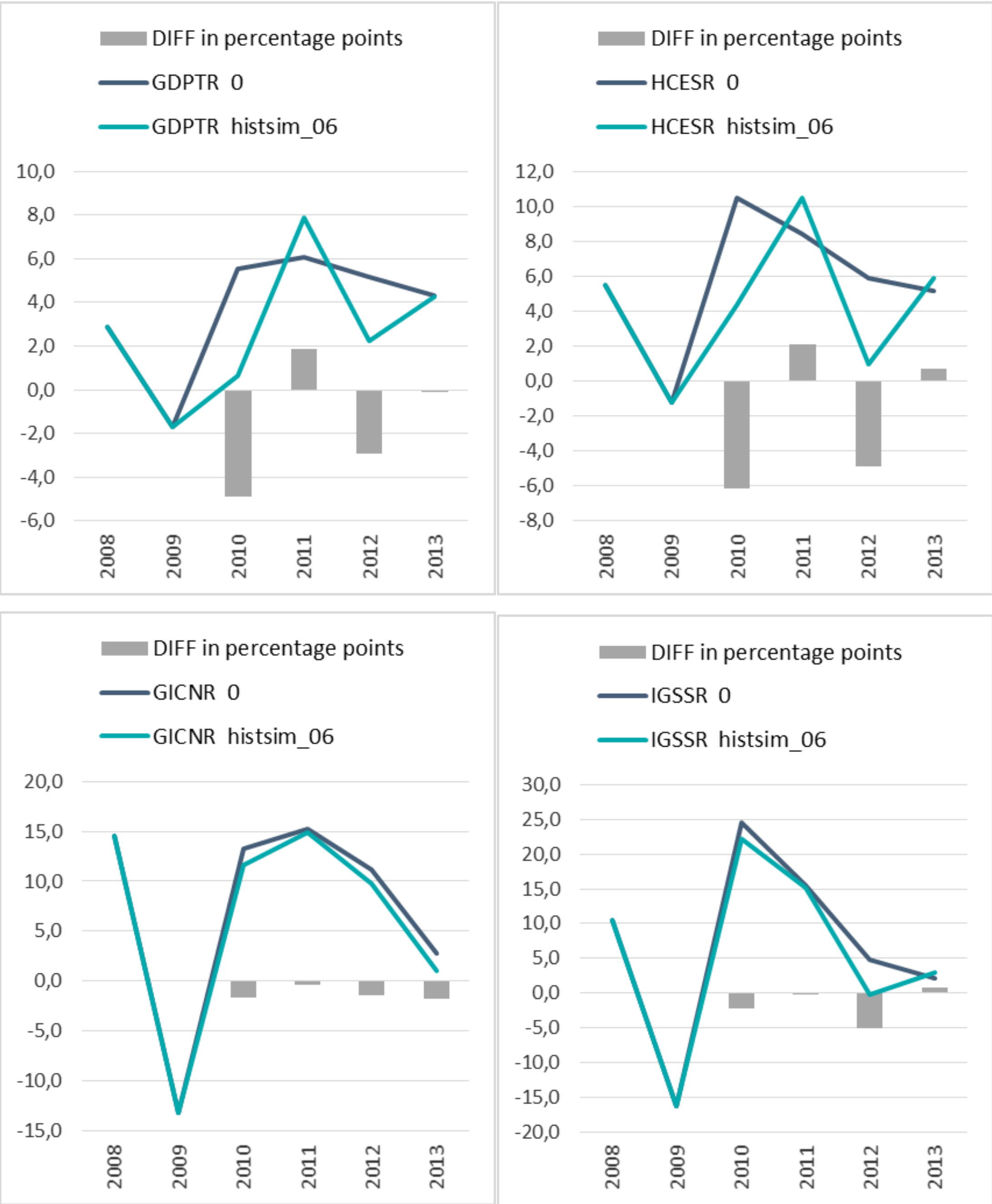
Table 2: Squared error measure

	Mean squared error (MSE)	Decomposition of MSE (controls for systematic under or overestimation)			
	$\frac{1}{T}\sum_{t=1}^{T}(\hat{x}-x)^2$	Proportion of forecast error (UM)	Regression proportion (UR)	Disturbance / residual proportion (UD)	UM+UR+UD
	Mean squared deviation of growth rates 2009-2013	the close to zero the better	the closer to zero the better	the closer to 1 the better	must sum up to 1
GDPTR	8,98	0,1638	0,2258	0,6105	1,0000
HCESR	16,50	0,1614	0,1467	0,6919	1,0000
CPONR	10,32	0,0190	0,3483	0,6327	1,0000
GCESR	0,00	0,000	0,0342	0,9658	1,0000
GICNR	1,97	0,5414	0,0553	0,4034	1,0000
EGGSR	15,98	0,0013	0,5929	0,4058	1,0000
IGSSR	7,72	0,2336	0,0132	0,7533	1,0000

Legend: GDPTR – real gross domestic product; HCESR – real private consumption; CPONR – real consumption of non-profit organisations; GCESR – real government consumption; GICNR – real gross investment; EGGSR – real exports; IGSSR – real imports

The swing in the positive and negative deviations of growth rates are shown for real GDP, private consumption, investments and imports in Figure 3.

Figure 3: **Differences in percentage points for real GDP (GDPTR), real private consumption (HCESR), real investments (GICNR) and real imports (IGSSR) between true (0) and projected (histsim_06) values**



The error measures of Table 1 and Table 2 have shown that the forecast is not super satisfactory. One reason may be that the forecast model is not good enough to forecast the Chilean economy. Some serious economic interlinkages may be missing. Another reason, however, could be also that the chosen time period is difficult to forecast. The world economic crisis in 2009 and its aftermath can give a first hint for this assumptions. Another is to compare the forecast with two alternative – naïve – forecasts. The results are shown in Table 3. For all GDP components, the result suggests that the applied forecast in better than the naïve forecasts.

The first naïve forecast uses the average growth rate between 2009 and 2013 as proxy for the average growth path of the Chilean economy. The relative comparison of both root mean squared errors indicates that “our” forecast beats the naïve forecast in all aspects. Especially real investments and real imports are much better forecasted by the COFORCE model than with the naïve forecast, with values close to zero. Real exports – instead – are pretty close to one, which indicates that both forecasts are more or less the same.

The second naïve forecast uses the growth rate of the previous year as proxy of this year’s growth rate. The COFORCE forecast is – again – better than the naïve forecast for all GDP components.

Because the COFORCE forecasts beat both two naïve forecasts, this indicates that the chosen time period for the ex-post simulation is a very tricky one. Naïve forecasts seem to be a less good choice to use, especially in times of economic distress. This is true, although the more sophisticated approach also has its deficits when looking at different error measure.

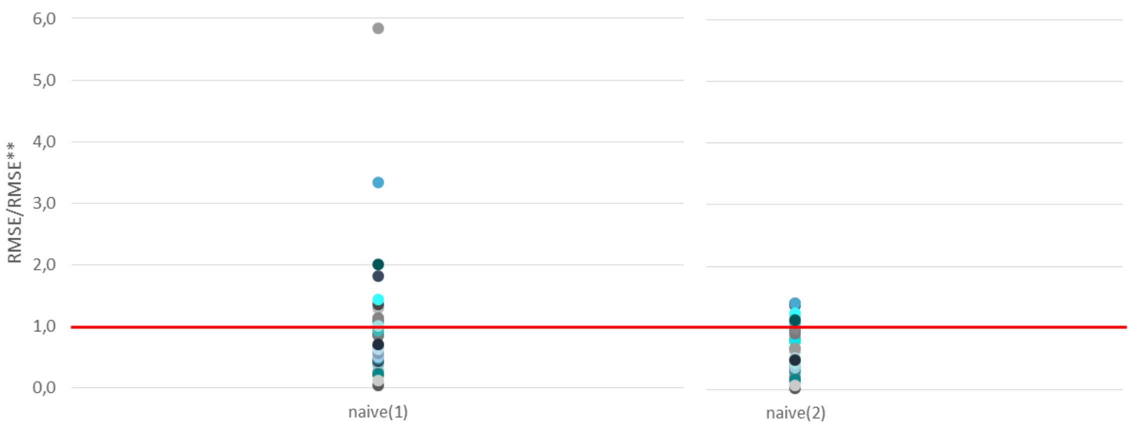
Table 3: Comparison with two naïve forecasts

	$RMSE/RMSE_{naive(1)}$	$RMSE/RMSE_{naive(2)}$
	Root mean squared error relative to naive(1) forecast	Root mean squared error relative to naive(2) forecast
	Naive(1): average growth rate 2009-2013	Naive(2): last observed growth rate t-1 used as forecast for t
	If value <1 forecast better than the naive forecast	
GDPTR	0,85	0,69
HCESR	0,84	0,58
CPONR	0,51	0,35
GCESR	0,00	0,00
GICNR	0,11	0,07
EGGSR	0,98	0,82
IGSSR	0,17	0,11

Legend: GDPTR – real gross domestic product; HCESR – real private consumption; CPONR – real consumption of non-profit organisations; GCESR – real government consumption; GICNR – real gross investment; EGGSR – real exports; IGSSR – real imports

All the above error measures and diagnostic checks can be calculated for all variables in the model COFORCE. However, the space is limited here to introduce them all. Still, the comparison with two naïve forecasts is shown for nominal production on sectoral level in Figure 4. With a view exceptions, the forecast is better (= below 1) for most of the sectors. However, for some sectors, the naïve forecasts remain the better choice. Especially the sectors *medical, dental and sanitation services* and *financial establishments* are better forecasted using a the naïve forecast (1). The naïve forecast (2) is generally less good. But also for naïve(2), some sectors seem to be better mirrored: for example *financial establishments* or *insurances*. Industry sectors are generally better captured with a more sophisticated modelling approach.

Figure 4 **Nominal production by 32 sectors – comparison with two naïve forecasts**



5 THE FINDINGS

This paper gave an introduction to the challenges and methods for conducting ex-post simulations as means of evaluating forecasts. It described the difficulties in interpreting ex-post simulations and introduced a wide range of standard error measure concepts that can be used for evaluating the robustness of a forecast.

The theoretical challenges have than been applied to the Chilean case. The forecast model COFORCE was configured in such a way to perform an ex-post forecast for the years 2009 to 2013.

The diagnostic check of the forecast has shown two things: the forecast is not super satisfactory, but it is better than naïve forecasts. As the ex-post evaluation was applied to an especially economically sensitive time, the results confirm that more sophisticated forecast models are especially useful in non-stable economic times.

REFERENCES

Andres, P. & Spiwoks, M. (2000) Prognosegütemaße. State of the art der statistischen Ex-post-Beurteilung von Prognosen. Sofia-Studien zur Institutionenanalyse Nr. 00-1.

Ash, J.C.K., Smyth, D.J. & Heravi, S.M. (1990) The accuracy of OECD forecasts of the interna-tional economy. Demand, output and prices. In: International Journal of Forecasting Vol 6: 379-392.

Baumgartner, J. (2002) Die Wirtschaftsprognosen von WIFO und IHS. Eine Analyse für die Achtiz-ger und Neunziger Jahre. WIFO Monatsbericht 11/2002. S. 701-741.

European Commission (2013) Evalsed Sourcebook: Method and Techniques. September 2013.

FISI & ECOFYS (2015) Electricity Costs of Energy Intensive Industries. An International Compari-son. For the German Ministry of Economic Affairs and Energy. July 2015.

Großmann, A., Lehr, U., Lutz, Ch., Mönnig, A. & Kleissl, S. (2016) Planning Policy Impact As-sessments and Choosing the Right Methods: manual for Development Practi-tioners. Deutsche Gesellschaft für International Zusammenarbeit (GIZ) mbH

Lutz, Ch.,Lindenberg, D. & Kemmler, A. (2014) Macroeconomic Effects of the Energy Transition. Final Report Summary. Project No. 31/33. For the German Federal Min-istry of Economic Affairs and Energy. September 2014.

Mönnig, A. & Bieritz, L. (2019): Copper Forecasting Chile – COFORCE Methodology Re-port. GWS Discussion Paper 2019/3, Osnabrück.

Woschnagg, E. & Cipan, J. (2004) Evaluating Forecasting Accuracy. UK Ökonometrische Prognose. University of Vienna. Department of Economics.