# A Vector Autoregressive Model for Forecasting Electricity Consumption in France

Stéphane Auray* Vincenzo Caponi†

October 15, 2020

**Abstract**

This paper provides a VARX approach for the estimation of electricity demand in France. Our methodology takes into account the complex relationship between weather variables and electricity demand, as well as the correlation between electricity and macroeconomic variables. We are able to provide a reliable conditional forecasting that, within the VAR framework, takes into account the common dependency of electricity consumption and other variables. While the VAR approach is not novel within this literature, our main contributions lie on the use of flexible functions that capture the role of weather to explain electricity consumption together with macroeconomic trend and cycle variables, and on the use of very detailed and comprehensive data on actual metered consumption of electricity in France. In-sample and out-sample forecasts provide evidence that our method is reliable for predicting future scenarios conditional on exogenous variables.

JEL: Q43; Q47

Keywords: Electricity Forecast;

*CREST-Ensai and ULCO. Email: stephane.auray@ensai.fr
†University of Sassari, CREST, IZA. Email: vcaponi@uniss.it

# 1 Introduction

Predicting electricity demand is vital in the energy industry. Forecasting models are used in the literature to predict electricity needs, prices and for predisposing the production of electricity to satisfy the amount of electricity demanded at certain times of the year or day or in different locations. It is also important to recognise that short and long-term demands of electricity (hours to short-term weeks and months to long-term years) respond to different variables, therefore a forecasting model should be flexible enough to capture this behavior. In the short term, weather variables are generally the main drivers of electricity consumption, however, even in the short term weather variables can have different effects when interacting with other economic variables, such as macro variables that capture the cyclically of economic activity, or trend variables that capture structural changes in the use of electricity for economic activity. As an example, we could expect that during hot days in summer electricity consumption increases because of the need of providing conditioned climates in many productive or residential places. However, we could also expect that during a downturn of the economy, this effect is reduced because of a reduced economic activity overall. In the longer term, on the one hand, the evolution of technologies that require less and less energy consumption, or that allow firms and families to produce their own, have a trend impact on the demand faced by traditional providers, on the other hand we also see more and more a shift from other means of energy to more electricity, for example the use of heat pumps instead of other traditional heating systems based on oil, or even the emerging market of electric cars.[1]

There are several time series models that are used in the electricity demand forecast-

[1]There exists an extensive literature on the elasticity of the electricity demand with respect of its price that also highlights this peculiar behavior in time. See among others Filippini (2011), Lim, Lim and Yoo (2014), and more recently for France Auray, Caponi and Ravel (Forthcoming 2020).

ing literature that depend on the availability of data and the range (time domain) of the forecast. AR(I)MA models are generally proposed when the data are relatively rich in terms of the number of observations of the endogenous variable, high frequency and long time span, but there are no or only few covariates that can explain and predict the behavior of that variable. These models are very reliable as long as there are not major changes in the underlying data generating process, that is, as long as the agents that interact in the electricity market do not change their behavior and that the main conditions that explain that behavior cyclically repeat themselves. Yet, even for the weather we cannot be sure about its regularly, much less we can trust about other fundamental variables such as technological changes, and other economic and demographic changes.

To address these issues and provide a more sound forecasting in the longer term, a small but significant number of studies offer models based on linear regressions that deal with a larger number of covariates. These models are usually used when the time domain of the study is longer and the forecast is also sought for a longer period, usually years. These models often include economic, weather and demographic variables to explain and predict aggregate energy demand. As an example, Bianco, Manca and Nardini (2009) propose a forecast model based on linear regressions that include macroeconomic variables such as GDP, GDP per capita and population between 1970 and 2007 and a forecast until 2020.[2]

The model we propose here is a natural extension of the two models above and, as such, gives us the opportunity on the one hand to model the behavior of the endogenous series taking into account its regularities in time, on the other hand we extend the model to include a series of covariates that the literature has so far suggested be important in predicting the consumption of electricity. Moreover, the VAR model we propose is

[2]For a more exhaustive review of different models of energy demand forecasting see Suganthi and Samuel (2012).

flexible enough not to impose causal relationships where it is not clear how the causality works. In our context this is an important issue as we would expect electricity to affect and be affected at the same time other macroeconomic variables. Our model will therefore predict all the variables that enter as endogenous based on their past history and the correlations among them. Yet, some variables are clearly exogenous to our model, in particular weather variables, hence we will take advantage of exogenous regressors using a VARX model.[3]

The use of VAR models for the forecast of electricity consumption is not new[4], however, this approach is very little exploited compared to other approaches that rely more on past information on the endogenous variable only or that limit the extent of covariates to weather variables. Our paper brings several contributions to the literature. First of all, our data on electricity consumption are very accurate and based on actual meter readings of all meters in France.[5] As such, we do not rely on approximations based on the production of electricity. This is a very important issue, because using the production of electricity as a proxy for consumption is very problematic in that the production is itself based on the expectation of what the demand would be. The difference between the production and the actual consumption of electricity is mostly wasted electricity, something that energy providers, and society as a whole, want to minimize. We also do not rely on surveys conducted on consumers or others actors in the electricity market, but the actual readings of the meters of all households and firms in France. On the modelling side, other than proposing a VARX, we also model the

[3]See Kaytez, Taplamacioglu, Cam and Hardalac (2015) for models that extend the multivariate linear regression approach to the use of artificial neural network. These models contrast to more traditional VARIMAX models in that they are capable to treat non-linearities more efficiently, but at the expense of economic interpretation, which makes them very sensitive to changes in the structure of the economic system in which the forecast needs to be used.

[4]See Ohtsuka and Kakamu (2013) for an example applied to Japan.

[5]Our data exclude French territories that are not within European geographical borders.

response of electricity demand on weather variables with more flexibility than commonly done and with a higher degree of refinement in terms of the variables used. That is, while it is common to use heating and cooling degree days (HDD and CDD), based on the deviation from "desirable" temperatures, as predictors of electricity demand, these methods implicitly assume a V effect of temperatures on consumption. To avoid imposing this structure we proceed by following a semiparametric approach to estimate the intra-month density of temperature for each month and approximate this density by a fourier transformation, we the use these approximations as variables for our Temperature Response Function (TRF).[6] This gives more flexibility and allows us to estimate rather than assume the threshold points of "desirability". Finally, we use all meteo station variables available instead than a weighted average, but because of the strong correlation among them we proceed to making a Principal Component Analysis (PCA) to reduce the variables to a reasonable and parsimonious subset.

For the rest of the paper, next section briefly describes in some details the econometric techniques that we use for our estimations, in particular the method to estimate the Temperature Response Function and the standard VARX model; section 3 the data and the pre-treatment before inputting the into the VARX model; section 4 presents the results of the VARX analysis; Section 5 concludes.

# 2 Estimation Methods

As we have discussed above, one main contribution of our paper is to use the Temperature Response Function within a VARX estimation to increase the flexibility of the model and therefore its predictive capacity. It is well known that the effects of temperature on electricity consumption is important and non linear and much of the effort in estimating

[6]Chang, Kim, Miller, Park and Park (2016)

and predicting electricity consumption is made in estimating the complex effect of this variable. Moreover, we very often have the possibility to observe temperatures at very high frequencies, hourly or even half-hourly, while much less often have electricity data with the same frequency domain. This is our case, where the consumption is measured monthly, while we have half-hourly data for temperatures. To avoid the strong assumptions made when using HDD and CDD measures of excess temperatures, as specified above, we resort to the method outlined below.

## 2.1 The Temperature Response Method

First of all, it is useful to think that the effect of temperature on electricity consumption is continuous and specified by a continuous function that we can call $g(r)$, where $r$ is a variable that measures time. We can think of temperature as a stochastic variable that follows some probability law that we can call $f(r)$ and therefore, assuming a multiplicative effect, we can define an average effect for a given interval in time as,

$$\tau_t = \int_{t0}^{t} f_t(r)g(r)dr \tag{1}$$

where we can think of the interval to be one month and the temperature distribution $f_t(r)$ specific to that month. We can also think of the variable $r$ in a discrete time domain as measuring fractions of the month, for example half-hourly fractions. Our problem is therefore to estimate the function $g(r)$ that we call the Temperature Responde Function. We start therefore from the following problem,

$$y_t = \tau_t + \epsilon_t = \int_{t0}^{t} f_t(r)g(r)dr + \epsilon_t \tag{2}$$

where $\epsilon_t$ is a mean zero error independent of $f_t$. Next step is to approximate the TRF

with a flexible Fourier function that takes care of different degrees of cyclicality,

$$g(s) = \sum_{i=0}^{p} c_i s^i + \sum_{j=0}^{q} [c_{1j} cos(2\pi js) + c_{2j} sin(2\pi js)] \tag{3}$$

finally, plugging equation (2) into equation (3) we obtain,

$$y_t = \sum_{i=0}^{p} c_i x_{it} + \sum_{j=0}^{q} [c_{1j} x_{1jt} + c_{2j} x_{2jt}] + \epsilon_t^{pq} \tag{4}$$

where $x_{it} = \int s^i f_t(s) ds$, $x_{1jt} = \int cos(2\pi js) f_t(s) ds$, $x_{2jt} = \int sin(2\pi js) f_t(s) ds$, and $\epsilon_t^{pq}$ tends to $\epsilon_t$ for $p$ and $q$ that go to infinity.[7]. In order to implement the method in practice, we need first to estimate the density functions $f_t$ of temperatures within all months for which we have electricity observations. This is achieved with the usual non-parametric kernel method.[8] Once we have an estimation of densities $\hat{f}_t$, we can integrate over the polynomial and trigonometric functions to obtain the $x's$ in equation (4) that we will interpret as regressors in a OLS model in order to then obtain an estimate of the coefficients $c's$. Finally, our TRF can be computed as,

$$\hat{g}(s) = \sum_{i=0}^{p} \hat{c}_i s^i + \sum_{j=0}^{q} [\hat{c}_{1j} cos(2\pi js) + \hat{c}_{2j} sin(2\pi js)] \tag{5}$$

The function in equation (5) tells us the response of electricity consumption to temperatures $s$. As we will see later in Figure (1), the proposed method is capable of estimating a complex non-linear relationship between temperatures and electricity that is important to take into account for precise forecasts. The next sub-section briefly introduces the VARX model we use. Within our model we slightly deviate from the

[7]For all the details on the method see Chang et al. (2016) and Chang, Kim, Miller, Park and Park (2014)

[8]See Li and Racine (2007) for an extensive guide on non-parametric and semi-parametric statistical methods.

TRF model presented above in the sense that we do not have a single OLS regression to estimate the $c's$ coefficients, rather the $x's$ of equation (4) enter as exogenous variables within the VARX model. Moreover, because we do not only have one single temperature for each period of time but 32 measured in different locations of the French territory, and because the measures are quite correlated among themselves, we resort in a Principle Component Analysis (PCA) to reduce the dimensionality of the exogenous temperature related variables taking into account only the first few that explain at least 95% of the variation of all $x's$.

## 2.2 The VAR Model

The vector autoregressive model is a extensively used tool in macroeconomics and, more in general, in time series analysis. The econometric model is very agnostic about the economic theory and very flexible at the same time, which makes it on the one hand no more than an explorative tool for economic modelling unless we impose more structure on it, on the other hand a very powerful forecasting tool. In the case of VARX models, forecasting is all more interesting as future values of the variable of interest can be obtained conditional on the realization of values of the exogenous variables. In our study we can deliver predictions of electricity consumption under alternative scenarios of weather forecast and other exogenous variables.

The VAR is the natural extension of the AR(p) model to a model in which we allow variables to be related one another. Let's take for example two variables $y_{1,t}$ and $y_{2,t}$ that are, for any possible reason, correlated and are also AR processes. In this case we have that,

$$y_{1,t} = \alpha^1_{10} + \alpha^1_{11} y_{1,t-1} + ... + \alpha^1_{1p} y_{1,t-p} + \alpha^1_{20} y_{2,t} + \alpha^1_{21} y_{2,t-1} + \alpha^1_{22} y_{2,t-2} + ... + \alpha^1_{2p} y_{2,t-p} + \epsilon_{1,t}$$

and

$$y_{2,t} = \alpha^2_{20} + \alpha^2_{21} y_{2,t-1} + ... + \alpha^2_{2p} y_{2,t-p} + \alpha^2_{10} y_{1,t} + \alpha^2_{11} y_{1,t-1} + \alpha^2_{12} y_{1,t-2} + ... + \alpha^2_{1p} y_{1,t-p} + \epsilon_{2,t}$$

The above two equations represent a structural VAR model, that is, a model in which it is specified the relationship between the two variables of interest and their own lags. This model, however, cannot be directly estimated as the endogenous variables $y_1$ and $y_2$ are both in the right and left side of the equations, i.e., the system is not identified. If the structural model is that of interest, as in many economic studies, then we need to impose further assumptions in order to be able to identify it. However, if predicting future values of our the endogenous variables is the purpose of the model, then we do not need to identify the parameters as they are in the above equations. To see why, let's look at the simpler case in which $p = 1$, then the structural VAR becomes,

$$\begin{aligned} y_{1,t} &= \alpha^1_{10} + \alpha^1_{11} y_{1,t-1} + \alpha^1_{20} y_{2,t} + \alpha^1_{21} y_{2,t-1} + \epsilon_{1,t} \\ y_{2,t} &= \alpha^2_{20} + \alpha^2_{21} y_{2,t-1} + \alpha^2_{10} y_{1,t} + \alpha^2_{11} y_{1,t-1} + \epsilon_{2,t} \end{aligned} \tag{6}$$

We can rewrite the system as follows,

$$\begin{bmatrix} 1 & \alpha_{20}^1 \\ \alpha_{10}^2 & 1 \end{bmatrix} \begin{bmatrix} y_{1,t} \\ y_{2,t} \end{bmatrix} = \begin{bmatrix} \alpha_{10}^1 \\ \alpha_{20}^2 \end{bmatrix} + \begin{bmatrix} \alpha_{11}^1 & \alpha_{21}^1 \\ \alpha_{21}^2 & \alpha_{11}^2 \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \end{bmatrix} + \begin{bmatrix} \epsilon_{1,t} \\ \epsilon_{2,t} \end{bmatrix}.$$

In matrix form we can write,

$$BY_t = \Gamma_0 + \Gamma_1 Y_{t-1} + \epsilon_t$$

where $B$ $\Gamma_0$ and $\Gamma_1$ are the matrices that collect the coefficients, $Y_t$ is the vector of variables $y_{1,t}, y_{2,t}$ and $Y_{t-1}$ the vector with the same variable and lagged values. This model can be transformed by inverting the $B$ matrix into,

$$Y_t = B^{-1}\Gamma_0 + B^{-1}\Gamma_1 Y_{t-1} + B^{-1}\epsilon_t$$

$$Y_t = \Pi_0 + \Pi_1 Y_{t-1} + \eta_t \tag{7}$$

The model in equation (7) is called the reduced form of the VAR and can be readily estimated by OLS equation by equation. All the coefficients in the $\Pi$ matrices are in fact identified, although, without additional assumptions, it is not possible to derive the original structure of the model from the reduced estimation. For forecasting, however, this is not a problem as all we need is in fact an estimate of the $\Pi$ matrices. Then, forecasting is done in a very similar manner as in the care of an AR() model, that is,

$$Y_{t+1}|t = \hat{\Pi_0} + \hat{\Pi_1} Y_t. \tag{8}$$

Notice here that the forecast is done for the whole vector of endogenous variables, which, in turns, increases the predicting power of the model. This is particularly true when we are interested in forecasting variables that are less predictable but which we

know are correlated with others that are more predictable.

Finally, an important extension of the VAR model, especially for our study, is to add exogenous variables that can explain our endogenous ones. The model is then very slightly more complicated by adding to the regressions a series of $X_t$ variables with or without lags. The general VARX model becomes then,

$$Y_t = \Pi_0 + \Pi_1 Y_{t-1} + ... + \Pi_p Y_{t-p} + \Theta_0 X_t + \Theta_1 X_{t-1} + ... + \Theta_q X_{t-q} + \eta_t \tag{9}$$

The VARX model can be very powerful and useful for forecasting as it uses more information to predict the future values of the variables of interest. In particular, values of the exogenous variables could be known in advance compared to our variable of interest, which would make the prediction more accurate. Exogenous variables can also be associated to policies, for example a change in tariffs etc..., in which case the VARX will produce reliable predictions on the effect of these changes.

The next section describes the data we use to implement the VARX model and, most importantly, the preliminary steps in treating those data.

# 3 Data Used and Pre-Treatment

## 3.1 Electricity Data

Our objective is to predict future values of total consumption of electricity in France. For this purpose we have available the monthly series of realized consumption from Jan. 1st 2010 to Feb. 1st 2018. This data was provided by Enedis. Together with this data Enedis also provided power subscription and number of sites of energy delivery, number of days within a month in which tariffs for TEMPO or EJP customers are more onerous (effacement), plus a series of calendar data such as the number of holidays and working

days in a month.[9] In addition to this data we collected data on economic activity in France, namely GDP, total consumption, exports, investments, as well as employment and unemployment, salaries etc...[10] We also have detailed temperatures observations for 32 meteo stations in France, at the frequency of every half hour, as well as a calculated weighted average for the whole France (realized temperatures) and normal temperatures for one whole year to be used in repetition for the same day and half-hour every year. A measure of nebulosity, realized and normalized for France at the same frequency as the temperatures is also available to us.[11] In the next three sub-sections we present the building blocks of what will enter in the VARX analysis. We presents three models with sequential improvements in order to highlight the contribution that each step in the data pre-treatment represents. At the end of the section Table 2 resumes the relevant statistics to evaluate comparatively the three models, in particular a measure of goodness of fit, given by the R squared since all models are eventually computed as OLS regressions and the square root of the sum of squares of deviations between actual data and out of sample forecasted values.

## 3.2 Model 1: Building on Existing Models

In order to better understand the ground basis of our modelling strategy and therefore our specific contribution in terms of accuracy of predictions, we start with a simple

[9] TEMPO and EJP are two types of tarification used to make the price of electricity higher during days in which there is higher demand. The tarification follows an algorithm that takes into account primarily the daily forecasted temperature for the whole France, other than the day of the week and the month of the year. The EJP tarification is not used for new customers any longer and currently affects a very limited market, the TEMPO tarification is still used and offers six different prices identifying three different type of days - blue (cheap), white and red(expensive) and further differentiating between night and day. TEMPO tariffs are offered only to consumers that have at least subscripted a delivery of up to 9 KVA, and it represents a limited market within France.

[10] See Appendix 1 for a detailed list of sources.

[11] "Normal" is defined as the average for that day and half-hour in the closest thirty year period, in this case Jan. 1st 1980- De. 31st 2010. The weights to reconstruct the national average from the 32 stations are provided by Enedis and are calculated taking into account electricity consumption.

model that is the starting point of many forecasting models of electricity consumption. This model is exclusively based on weather variables and in particular its core is the deviation from "normal" consumption as defined by what it would have been the consumption of electricity if the weather variables considered where at their normal levels as defined in the footnote above. Common models that explain electricity consumption in terms of weather variables (particularly temperatures), are based on the concept of heating and cooling degree days (HDD_R and CDD_R), obtained taking the sum of all the positive intra-day differences between a heating threshold and the realized temperatures (or the realized temperatures and the cooling threshold) within a month. The variables created in this way are then used as regressors to explain electricity consumption. In our first model we use these definitions applied to the average national temperature for France, with thresholds 25.79 and 18.99 for cooling and heating respectively, together with TEMPO and EJP days, calendar variables (Eff and Days) and a dummy for July 14th (Bastille). All the variables in such a model are exogenous, some are also deterministic (calendar days), while the weather variables are clearly uncertain for forecasting purposes.[12]

## 3.3 Model 2: Adding Flexibility

There are two assumptions in the above method that bring some limitation to our analysis that can be avoided. One first assumption is that the aggregate effect of the climate of each station is a constant proportion of the average climate effect. This is implicit in considering only the weighted climate variable rather than all the single stations. The other assumption is that temperatures have a V effect on consumption (although we

[12]The thresholds are chosen in order to minimize the RMSE of the out of sample forecasting errors. As this is our favourite measure for comparing different forecasting models, choosing these thresholds gives us the best possible fit for this model.

actually use a quadratic function, a little more flexible), this assumption is implicit in the construction of HDDs and CDDs. To avoid imposing this structure to our predictive model we proceed, as explaine above in section 2.1, by following Chang et al. (2016) and non-parametrically estimate the intra-month density of temperature for each month and approximate this density by a fourier transformation, we then use these approximations as variables for our Temperature Response Function (TRF). This gives more flexibility and allows us to estimate rather than assume the threshold points. Moreover, we also use all meteo station variables instead than a weighted average. However, climate variables of stations in the same country are likely to be very correlated and, therefore having a large number of them will most likely not add much information, for this reason by a Principal Component Analysis (PCA) we reduce as much as we can the dimensionality of our set of variables.

Before proceeding to the use of all meteo stations and the PCA analysis, in order to better appreciate the advantage of using the TRF method, we first apply it to the realized mean temperature of France, so that we can directly compare this method with our earlier method. As shown in section 2.1, the TRF method allows us to estimate also the response of electricity consumption to temperature as a flexible non-linear function, Figure 1 reports the estimated TRF for Model 2. The shape of the function reported in the figure tells us a couple of things that are relevant to our analysis, one is that indeed the relationship seems to be highly non-linear and that at low temperatures the impact tends to flatten. Another feature is a high degree of convexity around the minimum impact, at about 22 degrees, which suggests that lower temperatures starting already at around 20 degrees start to have a strong impact, as well as higher temperatures than 22/23 degrees. This contrasts with the hypothesis implicit in the use of HDD/CDD days, that there exists an area around 20/22 degrees that is relatively flatten with

near zero impact of temperatures on electricity consumption. This convexity, however, could be the results of a high variability of temperatures across the country rather or together with a high sensitivity of the demand of electricity around "comfortable" temperatures. For this reason we further extend the model to include such variability within our explanatory variables.

## 3.4 Model 3: Space-Differentiated Weather

Our last step in preparing the data, is the use of information coming from 32 meteo stations rather than taking some weighted average of them to construct a national measure, we resort to principle component analysis (PCA). The PCA is, however, done not on the series of temperatures, rather on the $x's$ variables computed as explained in section 3.1. That is, for each meteo station we take the time series of 140256 half-hourly observed temperatures, we then estimate a non-parametric kernel density for each station and each month available $\hat{f}_{kt}$, with $k$ for meteo station and $t$ for month, then we compute the following,

$$\begin{aligned}
\hat{x}_{it} &= \int s^i \hat{f}_t(s) ds \\
\hat{x}_{1jt} &= \int cos(2\pi js) \hat{f}_t(s) ds \\
\hat{x}_{2jt} &= \int sin(2\pi js) \hat{f}_t(s) ds
\end{aligned}$$

we do this for $i = 1, 2, 3$ and $j = 1, 2, 3$. We therefore end up with 32x9 series of explanatory variables of length 96 months. The PCA is then applied to each of the 9 variables, choosing the component that explain at least 95% of the variability across the 32 stations.

Figure (1) The Temperature Response Function

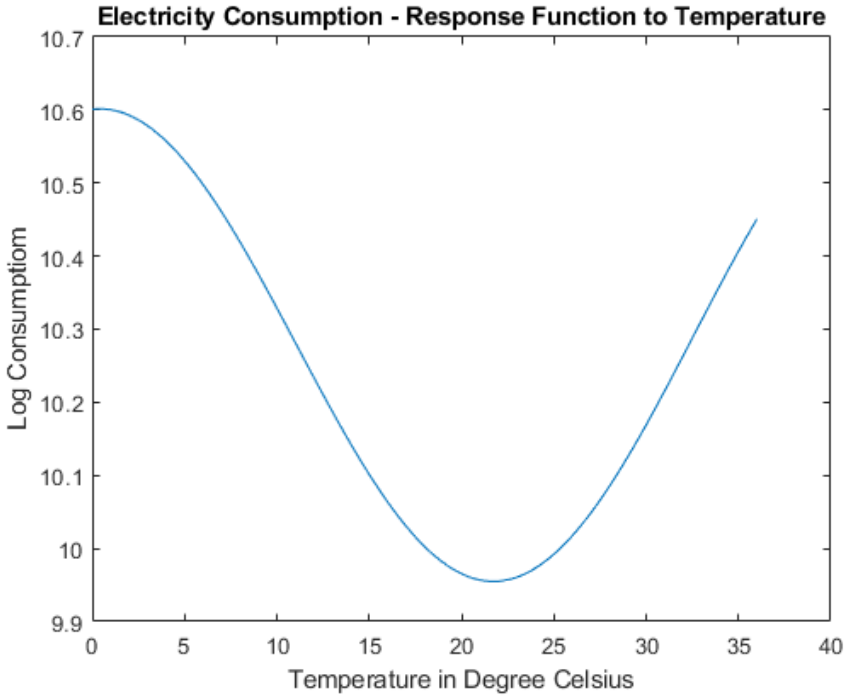



Table (1) PCA Explanatory Power of Components

| | $\hat{x}_{1t}$ | $\hat{x}_{2t}$ | $\hat{x}_{3t}$ | $\hat{x}_{11t}$ | $\hat{x}_{12t}$ | $\hat{x}_{13t}$ | $\hat{x}_{21t}$ | $\hat{x}_{22t}$ | $\hat{x}_{23t}$ |
|---|---|---|---|---|---|---|---|---|---|
| First Component | 98.6701 | 98.6954 | 98.6085 | 80.9078 | 74.4207 | 61.2123 | 98.5272 | 91.6324 | 72.4152 |
| Second Component | 0.5072 | 0.4905 | 0.5123 | 16.1372 | 21.7292 | 29.9643 | 0.5818 | 6.1188 | 20.7077 |
| Third Component | 0.3467 | 0.3696 | 0.4220 | 1.1179 | 1.5560 | 3.6862 | 0.4213 | 1.2264 | 3.5826 |
| Fourth Component | 0.1410 | 0.1367 | 0.1399 | 0.4446 | 0.5192 | 1.0991 | 0.1283 | 0.2222 | 0.6274 |

From Table (1) we can derive our choice of including only the first component for the first three $\hat{x}'s$, the first two components for $\hat{x}_{11}$, $\hat{x}_{12}$, $\hat{x}_{21}$ and $\hat{x}_2$ and the first three for the other two regressors.

## 3.5 Models' Comparison

Having explained how we built the variables in order to gain predictive power, we can look now at Table 2 to check the performance of the three models.

As we see in the table from the R-Squared the fit of the models improves from model 1 to 3. That is not surprising as from model 1 to model 2 we add some flexibility that allows for a better fit and in model 3 we also include more regressors. This better fit translates in a higher predictive power. All the measures of predictive power, that take into account the errors of out of the sample predictions, show improvements. The Root Mean Square Error, goes from 0.044 to 0.0364, while the Mean Absolute Error and the Mean Absolute Proportional Error go from 0.1926 to 0.1634, and from 0.0601 to 0.0511 respectively. Taking the our favorite measure, the RMSE, we can also see that the gain in predictive power of the second model compared to the first is around 8%, while for the third model is above 17%.

From this section we learn that adding flexibility in modeling the response of electricity consumption to temperature measures is very important in order to obtain reliable

Table (2) PCA Explanatory Power of Components

| | Model 1 | Model 2 | Model 3 |
|---|---|---|---|
| $R^2$ | 0.9475 | 0.9633 | 0.9676 |
| RMSE | 0.0440 | 0.0403 | 0.0364 |
| RMSE % gain | 1 | .0841 | 0.1727 |
| MAE | 0.1926 | 0.1801 | 0.1634 |
| MAPE | 0.0601 | 0.0562 | 0.0511 |

forecasts. In the next section we take this consideration into account to build a VARX model that reliably allows us to forecast electricity consumption taking into account not only exogenous but also endogenous variables.

# 4 VARX Analysis: Model 4

In this section we document our VARX forecast analysis. In particular we describe the variables we used as endogenous and as exogenous, we describe some details on the choice of lags and inferencing and discuss the forecast results.

We introduced three types of variables in our analysis: economic variables, weather variables and "technical" variables. Economic variables we believe are important as higher economic activity needs more energy, hence we postulate a correlation between electricity consumption and economic activity. Among the variables that describe economic activity we included GDP, total consumption of goods and services and total employment as share of working age population. All variables are used in natural log transformation. For "technical" variables, we indicate all other variables that have a relationship with electricity consumption but are not classified among the first two types. In particular we use the number of days in a month, the number of Saturdays or Sundays, and other variables such as the number of sites that deliver electivity etc... Among all these variables we divide between endogenous and exogenous. Endogenous variables are those that are correlated but may cause or be caused at the same time by electricity consumption, for example GDP we treat is as endogenous. The reason is that for producing more GDP there is the need of higher electricity consumption, therefore there is a causation from GDP to electricity, however, it may be that lower prices of electricity increase its consumption and at the same time GDP, in which case the causation is rather on the other sense. For VAR specification we don't need to take a stand on causation when

we are interested only on forecast, it is however important to recognize the possibility of erogeneity. We also include (the log of) wind energy production as an endogenous variable, for the same reason as above. As exogenous variables we include the number of days in a month, the number of Saturdays or Sundays as well as weather variables. In this case we are confident that they are indeed exogenous as, quite obviously, the number of days in a month does not change conditional on energy consumption, the same can be said for weather conditions, although there is increasing evidence that even at micro-climate levels energy consumption might have an effect on temperatures. We tried many different specifications with different sets of variables, while the results change a little, they do not dramatically when the choice is made wisely. We present only a sub set of results.

In order to avoid the problem of unit roots with the VAR system, and in order to treat seasonality, we proceed by taking seasonal differences of the time series, i.e. we use series that represent changes (given the log transformation, percent changes) of one month compared to the same month one year earlier. We loose 12 observation by differentiating, but this procedure is necessary to avoid meaningless results to due spurious correlation arising when integrated series (with unit roots) are used.

Here we present the graphs of all the endogenous variables in which we can see the model fit together with the predicted path for one year ahead. In this first specification we use as exogenous variables the temperature factor variables, as well as number of Sundays, Saturdays, number of days in the month, and tempo days red and white. For all these exogenous variables we have data until December 2017, while for the other variables used in the VARX estimation we have data only until March 2017. Therefore, the month from April to December 2017 are predicted. Indeed, we also have data for actual realization of consumption for those months, which we will compare with our

predications.

The following Figure shows the predicted results against the actual ones. In particular, the values on the right of the red vertical line of actual observations were not used to estimate the model, therefore, the difference between the actual and predicted results from April 2017 on can be taken as the forecast errors that were made. We also did not use the number of sites as well as the power subscripted actual values from April 2017, but their forecast as shown in the figures above.

Table 3 replicate in part the comparison table in the previous section adding the performance of the VARX model. While the MAE and MAPE do not show a significant improvement of the VARX compared to Model 3, the RMSE shows an additional improvement of more than 5% making the gain of using the VAR 22.41% compared to the original Model 1. This is an important results given also that the VARX model includes several variables that are endogenous and therefore forecasted as well for the out of sample forecast. Overall the VARX model suggests that not only is important to model flexibility properly, but also to conduct forecasts taking into account the complex relationship of electricity consumption with other macroeconomic variables.

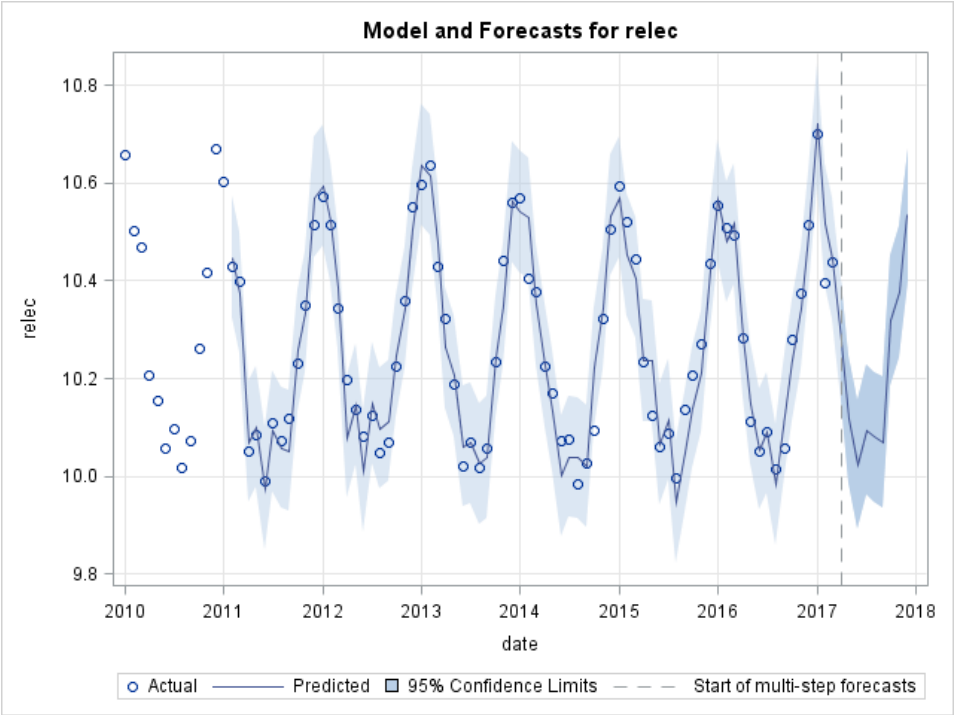
## 4.1 Diagnostic of the VARX model

The above model estimates a VARX(2,0), that is, two lags were chosen among the endogenous variables and none among the exogenous ones. When it comes to how to specify the model, i.e. the variables to include, which ones endogenous and which

Table (3) PCA Explanatory Power of Components

| | Model 1 | Model 2 | Model 3 | VARX Model |
|---|---|---|---|---|
| RMSE | 0.0440 | 0.0403 | 0.0364 | 0.0337 |
| RMSE % gain | 1 | .0841 | 0.1727 | .2241 |
| MAE | 0.1926 | 0.1801 | 0.1634 | 0.1635 |
| MAPE | 0.0601 | 0.0562 | 0.0511 | 0.0510 |

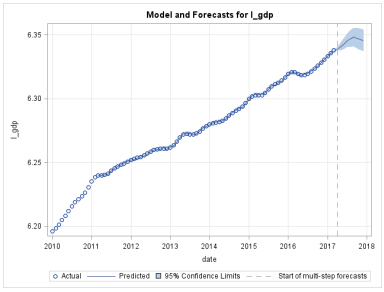
Figure (2) Forecast of Log Electricity Consumption



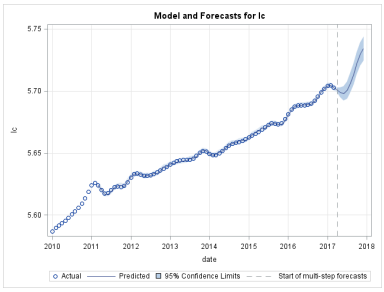

exogenous, and the number of lags, there are several factors that help choosing. First of all, economic and technical knowledge can guide in the choice of the variables to consider. We discussed this point above. However, in estimating the model we can include or exclude variables based on a series of tests on the significance they have in predicting our variables of interest. The same logic applies also to the choice of lags to include, on the one hand the more lags we include, the more we can explain in the model, on the other hand though, too many variables and lags can make the model difficult to interpret and less efficient in forecasting (i.e. increase the margin of the forecasting error). Table (4) shows the estimates of the autoregressive parameters in the model (for the exogenous parameters there are 7x24, too many to synthesize them here).

The table shows the autoregressive coefficients representing the matrices $\Pi_1$ and $\Pi_2$ of the VAR system. All coefficients have an associated standard error, but given the number of coefficients it is easier and more informative to conduct tests on groups of them as to

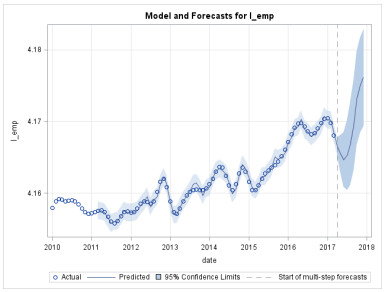
Figure (3) Forecast of Other Endogenous Variables



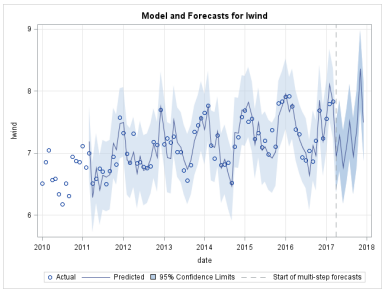

(a) Log of Real GDP



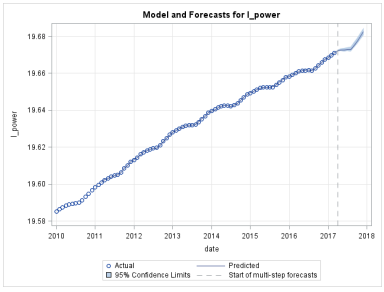

(b) Log of Consumption of Goods and Services



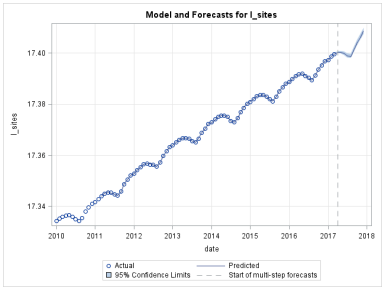

(c) Log of Employment Share




(d)




(e) Power Subscription




(f) Number of Sites

Figure (4)　Forecast vs Actual Total Electricity Consumption

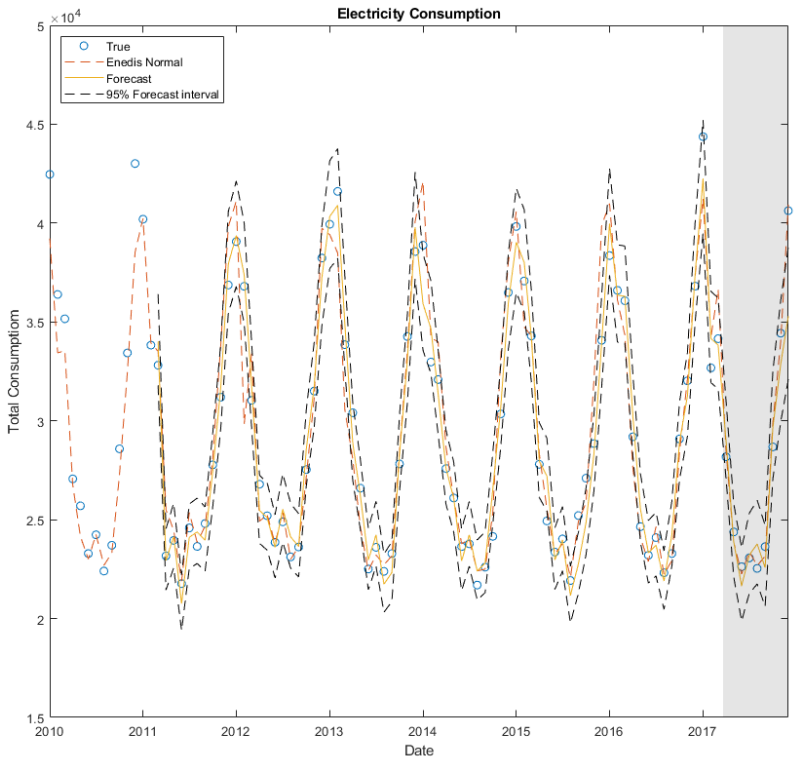



Table (4)　AR Coefficient Estimates

| Lag | Variable | relec | l_gdp | lc | l_emp | lwind | l_sites | l_power |
|---|---|---|---|---|---|---|---|---|
| 1 | relec | 0.0689 | 2.6082 | 14.8233 | 0.0467 | -26.2543 | 74.4038 | -37.5197 |
| | l_gdp | -0.0002 | 1.2921 | 0.0569 | 0.0004 | -0.4484 | -0.1974 | 0.4415 |
| | lc | -0.0052 | -0.3106 | 1.7117 | -0.0006 | -0.2971 | 1.609 | -1.6856 |
| | l_emp | 0.2905 | -30.4385 | 25.9606 | 0.141 | 55.0496 | -177.9324 | 269.5842 |
| | lwind | 0.0001 | 0.0667 | 0.0215 | 0.0002 | 1.7158 | 0.3063 | -0.6425 |
| | l_sites | 0.0013 | 0.0319 | -0.0078 | -0.0001 | 0.0329 | 0.7722 | 0.1085 |
| | l_power | -0.0002 | 0.0038 | 0.0096 | -0.0002 | 0.0158 | -0.1047 | 1.0288 |
| 2 | relec | 0.017 | -2.9569 | -13.7336 | 0.0657 | 28.3337 | -76.4171 | 39.8822 |
| | l_gdp | -0.0001 | -0.4638 | -0.0297 | 0 | 0.2855 | -0.266 | -0.2987 |
| | lc | -0.0059 | 0.3839 | -0.7736 | 0 | 0.0765 | 0.4463 | 0.4981 |
| | l_emp | 1.6399 | 54.8287 | -47.5257 | -0.0878 | -54.8917 | 128.4344 | -246.2731 |
| | lwind | 0.0009 | -0.0899 | 0.0032 | 0.0006 | -0.8948 | -0.2911 | 0.5669 |
| | l_sites | -0.0007 | -0.0164 | 0.0186 | -0.0002 | -0.017 | 0.0498 | -0.003 |
| | l_power | -0.0015 | 0.0108 | 0.0213 | -0.0001 | -0.0384 | 0.3376 | -0.1401 |

test the significance of various parts of the model. In particular we test the significance of all the exogenous variables together, with a Wald test, in their contemporaneous effect and in their lagged effect. Moreover, we also test the effect of each endogenous variable on all the others (excluding on its own). Table (5) resumes the battery of tests we do.

The Table shows interesting results. First of all, the set of exogenous variables are very significant in explaining electricity consumption. To the extent that these variables can be known, or well predicted, in advance (in particular weather variables) this gives more confidence to the forecast power of the model. Among the endogenous variables that seem strongly correlated with electricity consumption are especially consumption of goods and services and the share of employment; GDP, wind electricity production as well as number of sites and power subscribed do not seem very significant. However, when we turn to the whole system, GDP and wind production become also significant while consumption it is not, sites and power remain not significant. This suggests that while consumption has an important direct effect on electricity, it is itself explained by other variables, GDP for example. Therefore, in order to provide a good forecast of electricity consumption, we also need to include other variables such as GDP that can help us to provide a good forecast of consumption. This is, indeed, one of the main strengths of the VAR estimation in forecasting as we can easily include variables that can also have an effect through other variables, without necessarily modelling the causal relationship of these effects.

# 5 Conclusion

We proposed a VARX approach for the estimation and forecasting of the demand of electricity in metropolitan France. We paid special attention to the treatment of weather variables as we know that, in the short to medium run, they are highly correlated to

Table (5) Testing on Parameters Significance

| Test | DF | Chi-Square | Pr > ChiSq |
|---|---|---|---|
| All Exogenous Variables on Elec. (Lag 0) | 24 | 37.39 | 0.04 |
| Log GDP on Elec. | 2 | 0.61 | 0.7381 |
| Log Cons. on Elec. | 2 | 8.2 | 0.0166 |
| Log Emp. on Elec. | 2 | 7.94 | 0.0189 |
| Log Wind on Elec. | 2 | 6.51 | 0.0386 |
| Log Sites on Elec. | 2 | 3.45 | 0.1781 |
| Log Power on Elec. | 2 | 1.47 | 0.4792 |
| Log GDP on System | 12 | 26.69 | 0.0086 |
| Log Cons. on System | 12 | 20.81 | 0.0533 |
| Log Emp. on System | 12 | 29.69 | 0.0031 |
| Log Wind on System | 12 | 29.65 | 0.0032 |
| Log Sites on System | 12 | 22.29 | 0.0344 |
| Log Sites on System | 12 | 25.4 | 0.013 |

electricity consumption. We showed how relying on a more flexible estimates of the effects of weather variables we can enhance greatly the predicting power of our model. We then turn to the a VARX model, which includes demographic and economic variables, in order to account for complex relationships between the consumption of electricity and other macroeconomic variables that are endogenous to electricity itself. We showed that our VARX model has the property to perform very reasonably not only in the short run, but also in the longer run providing out-of-sample forecasting that are reasonably close to realized data. How model significantly contributes to the literature of energy forecasting suggesting a series of steps and an overall model capable of improving the predicting power of out of sample forecast. There are several improvements and checks that we were not able to perform in this study due to data limitation, particularly due to the limited time-span. We suspect that the VARX model would perform much better if we could run it with a longer time span as the VAR structure is more suited to take into account long run relationships. In this sense we would have like also to look at possible error correction mechanisms, which due to data limitation we could not assess. Another

issue that arise with our methodology is how to conduct forecasting when the exogenous variables are unknown in advance, a point particularly important for weather variables. The literatures proposes alternative solutions to this issue, one solution is to look at "normal conditions", that is for example average temperatures over the past 10 to 30 years; another solution is to forecast independently the series of exogenous variables, for example running AR(I)MA models, or, even better, to feed into the model the weather forecasts done by meteorologists. We believe that addressing these further issues within our proposed model will further improve our predicting capabilities.

# References


**Auray, Stephane, Vincenzo Caponi, and Benoit Ravel**, "Price Elasticity of Electricity Demand in France," *Economics and Statistics*, Forthcoming 2020.

**Bianco, Vincenzo, Oronzio Manca, and Sergio Nardini**, "Electricity consumption forecasting in Italy using linear regression models," *Energy*, 2009, *34* (9), 1413 – 1421.

**Chang, Yoosoon, Chang Kim, James Miller, Joon Park, and Sungkeun Park**, "A New Approach to Modeling the Effects of Temperature Fluctuations on Monthly Electricity Demand," *Energy Economics*, 09 2016, *60*.

___ **, Chang Sik Kim, J. Miller, Joon Y. Park, and Sungkeun Park**, "Time-varying Long-run Income and Output Elasticities of Electricity Demand with an Application to Korea," *Energy Economics*, 2014, *46* (C), 334–347.

**Filippini, Massimo**, "Short- and long-run time-of-use price elasticities in Swiss residential electricity demand," *Energy Policy*, 2011, *39* (10), 5811 – 5817. Sustainability of biofuels.

**Kaytez, Fazil, M. Cengiz Taplamacioglu, Ertugrul Cam, and Firat Hardalac**, "Forecasting electricity consumption: A comparison of regression analysis, neural networks and least squares support vector machines," *International Journal of Electrical Power & Energy Systems*, 2015, *67*, 431 – 438.

**Li, Qi and Jeffrey Scott Racine**, *Nonparametric econometrics: theory and practice*, Princeton University Press, 2007.

**Lim, Kyoung-Min, Seul-Ye Lim, and Seung-Hoon Yoo**, "Short- and long-run elasticities of electricity demand in the Korean service sector," *Energy Policy*, 2014, *67*, 517 – 521.

**Ohtsuka, Yoshihiro and Kazuhiko Kakamu**, "Space-Time Model versus VAR Model: Forecasting Electricity demand in Japan," *Journal of Forecasting*, 01 2013, *32*.

**Suganthi, L. and Anand A. Samuel**, "Energy models for demand forecasting—A review," *Renewable and Sustainable Energy Reviews*, 2012, *16* (2), 1223 – 1240.

# 6 Appendix A - Data Sources

As mentioned in the text above, the main source of our data is Enedis, that has provided observations for consumption of electricity, number of holidays, Sundays and Saturdays, within holidays days established as holidays as well (ponts), and the number of days in a month. It also provided the tariffs for Tempo and EJP customers. Temperatures were

also taken from Enedis. Other variables we used are total Employment as a fraction of working age population, from INSEE Wind electricity production from "Pègase" of the Ministry of Sustainable Growth and for the aggregate economic series, such as GDP, the IMF. Below Table (6) resumes the sources.

Table (6) Data Used

| Source | Name from Source | Name Used |
|---|---|---|
| IMF - National Accounts, Current Prices, Non-Seasonally Adjusted | | |
| Gross Domestic Product, Nominal, Domestic Currency | NGDP_XDC | GDP_N |
| Household Consumption Expenditure, incl. NPISHs, Nominal, Domestic Currency | NCP_XDC | Consumption as share of GDP_N |
| Exports of Goods and Services, Nominal, Domestic Currency | NX_XDC | Exports as share of GDP_N |
| Imports of Goods and Services, Nominal, Domestic Currency | NM_XDC | Imports as share of GDP_N |
| Government Consumption Expenditure, Nominal, Domestic Currency | NCGG_XDC | Government Consumption as share of GDP_N |
| Gross Fixed Capital Formation, Nominal, Domestic Currency | NFI_XDC | Investments as share of GDP_N |
| Change in Inventories, Nominal, Domestic Currency | NINV_XDC | Investments (part of) as share of GDP_N |
| IMF - National Accounts, Constant Prices, Non-Seasonally Adjusted | | |
| Gross Domestic Product, Volume | NGDP_R_K_IX | GDP |
| Pégase - Èlectricitè, approvisionnement et consommation en France, en GWh - Production brute d'èlectricitè èolienne (en GWh) | | Wind |
| INSEE - Personnes en emploi (taux d'emploi) au sens du BIT - Ensemble des 15 à 64 ans (001688428) | | Emp. |