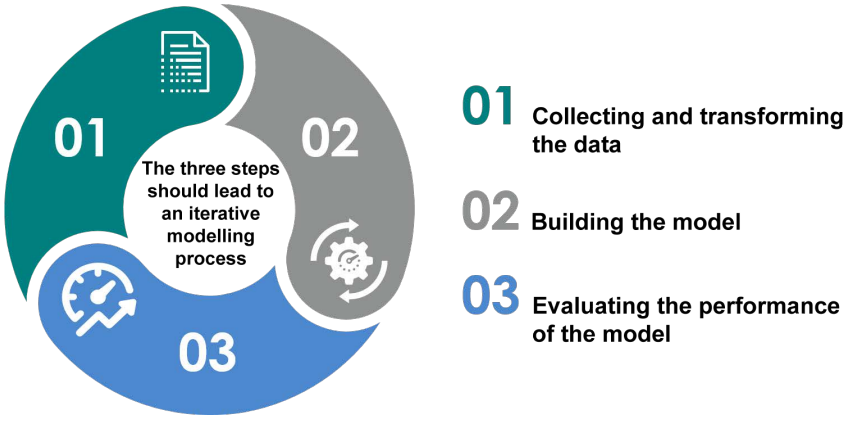
1. THE STATISTICAL APPROACH IS AT THE HEART OF CREATING FORECAST MODELS

Statistical models enable the use of an activity's historical performance measures to define the mathematical relationship between such measures and external variables (macroeconomic variables, banking market data, etc.) or internal variables (seasonality). The forecasts therefore consist of multivariate regression models, be they linear or non-linear[3].

In this context, a compromise must often be found between the ease of implementation and appropriation by the business lines, on one hand, and predictive power and statistical robustness, on the other hand. In order to arrive at this compromise, it is necessary to implement an iterative process in three steps (as detailed in figure 1):

1. Collecting and transforming the data
2. Building the model
3. Evaluating the performance of the model.

Figure 1: Iterative process leading to the creation of a model



01

- **Organisational implementation facilitating the collection of quality data.**
- **Transformation of data to make it usable** via the deletion of exceptional events, the completion of missing data, a seasonality adjustment and other potential mathematical transformations.
- **The choice of model in step 2 depends on the quality and historical depth of the data.**

02

- **Choice of model applied –** linear initially then more advanced methods are used depending on the data and the type of business studied.
- **Machine learning methods enable the orientation of the choice of variables; however, we will prefer more standard methods:**
  a) Easier to put in place
  b) Enables us to obtain results that are simpler to audit ("black box" issue of forecasts in Machine Learning).

03

- Use of **statistical tests** to check the mathematical robustness of the model.
- **Back-testing** of model over historical periods.
- **Sensitivity test** of model based on a shock to explanatory variables.

---

[3] Regression models are used to explain the evolution of a variable according to one (univariate model) or several variables (multivariate model). These regression models can be linear if there is a relationship of direct proportionality between the explained variable and the explanatory variables.

2. MANAGEMENT OF THE DATA: A PREREQUISITE FOR THE CONSTRUCTION OF STABLE MODELS

The quality of a statistical approach depends largely on the data on which the modelling works are based. Indeed, if the historical data from which the relationships with macroeconomic indicators (or other indicators) are defined include "polluting" effects, the models are less precise and can even lead to erroneous conclusions. In this context, the validation of the quality of the data by the business lines is a prerequisite for any statistical analysis. It is important to involve members of the business line teams in these considerations to capture all 'non-standard' items.

Once these verifications are complete, the data series can be transformed to improve the quality of the models to test in step (2). These transformations should take into account different types of effects, including:

- the representativeness of **exceptional business items** (one-offs), such as large mergers and acquisitions (jumbo deals) or fiscal shocks, which can reoccur in reality;
- the completion of **missing and/or asynchronous data**. For example, we might want to forecast an indicator on a monthly basis from explanatory variables that are only available quarterly. In this case, it is possible (i) to rely on the quarterly series (implying the loss of points of analysis), (ii) to interpolate on a monthly basis (linearly or not), or (iii) even to realise a filtering[4] by completing the data using statistically coherent estimations from indicators with a greater frequency;
- the **seasonality[5] to be adjusted**, for example in the case of non-seasonal explanatory variables. Seasonality transformations can be based on the ARIMA-12[6] algorithm or the STL process based on local regressions[7];
- the **smoothing of the series** to be explained, for example by calculating a moving average[8];
- the **introduction of a delay effect**, or lag, to the series to be explained.

Alternatively, other mathematical transformations can be applied to the series to iteratively improve the results of the evaluated models during the model evaluation in step (3):

- **data differentiation** – of the order 1 or 2 – with a short frequency (one quarter, for example) or a long frequency (one year, for example). This generally enables the correction of non-stationarity[9] statistical bias but can sometimes be unstable in projections;
- the application of functions (growth rate, square, logarithm) aiming to **capture non-linear effects**;
- the **use of models in co-integration** (see below) in case of the non-stationarity of the variables to be explained.

---

[4] This can be implemented using a Kalman filter, for example.

[5] Or more generally the autocorrelation of the series, which amounts to introducing an endogenous variable into the model.

[6] The ARIMA-X12 algorithm is a popular method of seasonality adjustment developed by the US Census Bureau. This method applies to series with monthly or quarterly seasonality. It is implemented in most statistical software and is one of the methods advocated by the European Statistical System (ESS).

[7] The STL ("Seasonal and Trend Decomposition Using Loess") procedure is a method of breaking down a time series into a seasonal component, a trend and residuals. As such, it is also a method of adjusting the seasonality that may be preferred in some cases to ARIMA-X12-type methods (especially in case of fluctuating seasonal components or in the presence of outliers).

[8] More generally, this can be incorporated, with the seasonality, into an ARMA-type (Auto Regressive Moving Average), ARIMA-type (AutoRegressive Integrated Moving Average) or SARIMA-type (Seasonal ARIMA) modelling process.

[9] The stationary (or not) character of a time series refers to the homogeneity of its statistical distribution over time. A weaker property used in practice (weak stationarity) is the fact of having its first two moments (mean and variance) constant, as well as an invariant autocorrelation function by translation over time.

The appreciation of the quality of the data collected as well as the characteristics of the transformations applied to them during the data collection in step (1) must be taken into account when choosing the model to develop during the modelling in step (2). Indeed, the historical depth of the data must be sufficient to capture distinct scenarios (crises, differentiated interest rate scenarios, etc.). Moreover, the data must present comparable business realities (for example, a scale effect linked to the number of traders on a desk or a reorganisation of the activity must be taken into account to harmonise the series from a statistical perspective).

3. ADAPTING THE CHOICE OF THE EXPLANATORY VARIABLES AND THE EXPRESSION OF THE MODEL TO THE ENVIRONMENT

Choice of model

The expression of the model must adapt to the needs of the users of the tool.

- Linear approaches, for example, are simpler to implement but do not allow the model to capture more complex relationships than affine relationships between the explained variable (or its growth) and the explanatory variables (or their growth). Coupled with the use of the non-linear transformations on the explanatory variables of the model, however, simple linear approaches enable the model to capture non-linearities. For example, the logarithm of mortgages can be correlated to the logarithm of the growth of household income. The use of logarithmic transformation enables the model to link the variables whose orders of magnitude are different.
- Machine learning methods[10], such as the random forest[11], are very good tools to orientate the choice of variables. They are, however, rarely retained because they are often complex to implement and difficult to audit for a regulator. Furthermore, they do not highlight exogenous drivers of activity and can remain too centred on autoregressive forms[12].

Models in co-integration

In case of non-stationarity[13], classic statistical models are unstable and specific techniques must be used. One central notion today is that of the model in co-integration for macroeconomic variables. A set of variables is co-integrated with the observed series if there exists a combination of variables that enables the cancellation of 'the stochastic trend' of the observed series to end up with a stationary series. For example, it has been demonstrated that in the United States, actual consumption per inhabitant and actual available income per inhabitant are co-integrated, highlighting a stable relationship between these two non-stationary series.

[10] These methods are part of what is now called 'Machine Learning', which aims to leverage data to determine the form of the model to be adopted, rather than specifying it upstream. These methods are based on the statistical analysis of a large number of data of various natures.

[11] Random forests are a family of machine learning algorithms that rely on sets of decision trees. The interest of this method is to train a set of decision trees on subsets of the initial dataset and thus to limit the problem of over-learning. This type of algorithm makes it possible to perform classification (estimation of discrete variables) and regression (estimation of continuous variables).

[12] An autoregressive model is one in which a variable is explained by its past values rather than by other variables.

[13] A random process is considered stationary if it is stable over time. Mathematically, this results in particular in a constant expectation (there is no trend) and a constant variance.

These co-integrated variables are therefore linked to the observed series by a 'long-term' linear equation, which can be interpreted as a macroeconomic equilibrium in relation to which the differences constitute temporary fluctuations. By looking at the previous example, a temporary fluctuation in consumption in relation to available income can occur in a given quarter, but it will have a comparably opposite effect on the future consumption of the next quarter, which tends to bring the two series towards their point of equilibrium represented by the long-term relationship.

The historical approaches to understand this type of relationship are those of Engle and Granger[14] or Johannsen[15], as well as the models called Autoregressive Distributed Lag (ARDL). All these models capture both long-term relationships and deviations from these equilibria via mean-reverting and error-correction models.

Choice of variables

Initially, the explanatory variables will be chosen from among all the transformed variables, thanks to simple correlation studies. The choice of explanatory variables can also be informed by the business line expertise, systematic classification approaches (such as principal component analysis[16]) or even their significance in machine learning methods (we then preserve the variables highlighted by the method but by applying classic statistical models to them).

On the contrary, certain variables will be excluded a posteriori by the statistical tests in step (3). In particular, too many variables can result in over-fitting and collinear variables to unstable regression coefficients[17].

Calibration of parameters

The method to estimate the parameters of the regression depends on the tests undertaken in step (3). They will be estimated either by the least squares estimator or, for example, by Yule-Walker[18] estimators to avoid the bias that is inherent to the existence of autocorrelation of residuals in the series used.

The problems raised by non-stationarity also concern the inference of the parameters of the estimated model, for which the usual asymptotic laws derived in the context of stationary series can lead to inconsistencies if used as such.

Notably, p-values (see below) and confidence intervals are no longer reliable in the context of non-stationary series or co-integration.

---

[14] Co-Integration and Error Correction: Representation, Estimation, and Testing, Robert F. Engle and C. W. J. Granger, 1987).

[15] Estimation and Hypothesis Testing of Cointegration Vectors in Gaussian Vector Autoregressive Models, Johansen, Søren, 1991.

[16] Principal Component Analysis (PCA) is a method of data analysis, which consists in transforming variables that correlate amongst themselves into new variables that are de-correlated from each other on the basis of their mathematical characteristics (orthogonal decomposition to own values).

[17] The Lasso or Ridge regressions allow the regularisation of the problem and the selection of variables of greater interest by introducing penalty terms.

[18] The Yule-Walker equations establish a direct correspondence between the parameters of the model and its autocovariances. They are useful for determining the autocorrelation function or estimating the parameters of a model.

4. EVALUATING THE MODELS GIVES CREDIBILITY TO THE STATISTICAL PROJECTION WORKS

The predictive power of the model must be verified by a set of tests. Statistical tests or backtesting can be done to support the choice of the model, even though neither of them is eliminatory as regards the choice of the model. We note that the verification requirement of these tests is to be weighted by the quality of the available data. The sensitivity of the model to a shock to the explanatory variables must be appreciated in all cases.

Statistical tests

The calculation of the significance of the variables (p-value) is important, but the estimation of the parameters and the calculation of the p-values must be corrected in case of the non-compliance of the basic assumptions of the linear regression[19]:

- **Stationarity of the time series[20]** (homogeneity of their distribution over time): the results of the linear regressions can be unstable over time if the series are not stationary, even in the case of a good R2. In this case, it is preferable to transform the variables (step (1)) or to choose a co-integration model (step (2)).
- **Homoscedastic residuals[21]** (constant variance over time) and/or more generally not self-correlated[22]: In case of non-compliance, this may be indicative of an unfound explanatory variable. This can significantly bias the variances and the confidence intervals of the coefficients. It is therefore necessary to correct the coefficients[23] or to modify the estimators used[24].
- **Normality of residuals[25]:** this assumption of linear regression is, however, rarely verifiable on small samples (asymptotic property) and is not necessary for the convergence of the parameter estimators.

Backtest (or cross-validation or out-of-sample performance)

If the historical depth permits, it is possible to measure the difference between the actual historical series and the model calibrated on a different time period. By repeating the exercise over several sub-periods, it is possible to verify the stability of the coefficients of the regression. An equivalent average error between the period tested and the calibration period is a good indicator that the model is not over-calibrated (over-fitted).

5. DIFFICULTIES

Modelling all the financial aggregates of a bank requires modelling activities of diverse natures relying on heterogeneous statistical models. In this context, the team responsible for the development of the models must adapt to the reality of each of the segments of activity.

This results in a plethora of statistical models to be articulated on a flexible platform enabling them to be linked to each other and to the data sources (business line databases notably) to propose results that are directly usable by the platform's teams.

---

[19] When the assumptions that provide the asymptotic distributions or the confidence intervals of the estimators are no longer satisfied, the confidence intervals can still be calculated by simulation (bootstrapping or resampling).

[20] The classic tests to run are those of Dickey-Fuller (increased), Phillips-Perron or Kwiatkowski-Phillips-Schmidt-Shin.

[21] Tests of Breusch-Pagan and of Goldfeld and Quandt.

[22] Tests of Durbin-Watson and of Breusch-Godfrey.

[23] Transformation of Yule-Walker (Cochrane-Orcutt generalised).

[24] Correction of the covariance matrix by the Newey-West estimator.

[25] Shapiro-Wilk test, for example.

Four types of difficulty must be overcome to constitute a sufficiently solid base of models to integrate to the platform:

- Difficulty in finding predictive statistical models on certain perimeters: not all activities are able to be modelled using a statistical approach, and some are more complex to understand (specific commissions, general fees, etc.). Moreover, the majority of classic statistical models struggle to capture non-linearities in past behaviours.
- Adding overlapping distinct effects to the activities' core modelling: Foreign exchange effects, concentration of portfolios, etc. Contagion effects, reputation effects and all feedback effects are particularly complex to capture.
- Difficulty in collecting quality data that is easy to update after the first modelling exercise: lack of depth of the data, homogeneity issues, etc.
- Organisational difficulties and issues with the tool.

6. CONCLUSION

Whilst banks now have recognised quantitative modelling teams, these skills are mainly concentrated in the Risk teams on credit risk and market risk issues. For most banks, prospective modelling based on statistical methods implies the constitution of specialist teams.

The methods discussed above provide a global vision of the statistical measures available to planning teams to build their projection models. The human dimension and the ability to recruit talent capable of building complex models is at the heart of the issue.

The development of a tactical approach, via agile tools, enables banks to create initial support for the platform and to distinguish the construction of the models from their industrialisation in the bank's systems.