Fostering European SMEs' internationalization using big data: the BIZMAP application¹

Jean-Noel Kien, Etienne Kintzler and Theo Nicolas,
Bank of France

¹ This presentation was prepared for the Workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the central banks and other institutions represented at the event.
Fostering European SMEs’ internationalization Using Big Data: 
The BIZMAP Application

Jean-Noël Kien, Etienne Kintzler, Théo Nicolas
Banque de France. E-mail: Jean-Noel.KIEN@banque-france.fr; Etienne.KINTZLER@banque-france.fr; Theo.NICOLAS@banque-france.fr. Address: 37 Rue du Louvre 75002 Paris.

Abstract

This paper proposes a decision-making tool (BIZMAP) that enables European small and medium-sized enterprises (SMEs) to visualize the most economically attractive European regions for the internationalization of their business activities. Building on more than 80 variables coming from seven different open access databases, we take advantage of big data and machine learning methods to include the most relevant ones in a standard gravity model of trade. In the end, we implement an interactive data visualization tool inside our BIZMAP application. Depending on the sector and the home country, we provide SMEs with a ranking of most promising European countries. Importantly, BIZMAP not only enables SMEs to understand what are the main drivers of this score but also offers the possibility to compare the 281 European regions with each other. Hence, by reducing information uncertainty abroad, BIZMAP is likely to improve the SMEs’ analysis of new markets through the visualization of harmonized territorial attractiveness indicators.

Keywords: SMEs, Trade, FDI, Big Data.


The views expressed in this paper are those of the authors and do not necessarily coincide with those of the Banque de France or the Eurosystem. The authors are grateful to Jérôme Coffinet, Jean-Brieux Delbos, Louis-Marie Harpedanne, Guillaume Gaulier and Antoine Berthou.
Non-technical summary

The internationalisation of economic activities opens up news opportunities for SMEs. However, some obstacles to their exploitation remain. Among them, the information deficit turns out to be one of the most salient. To tackle this issue, the BIZMAP application offers a decision-making tool that enables SMEs to identify the most economically attractive EU countries or regions for their internationalisation (exports or foreign direct investments – FDI).

The principle of the application is straightforward: after filling out all the necessary fields (sector, home country and type of internationalisation), BIZMAP first provides the SME manager with an interactive visualisation of the most promising national markets ranked by scores. The latter are based on a wide range of criteria aggregated into a five-dimension indicator: economic perspectives, standard of living, infrastructure, financial conditions and institutional environment. Then the application goes even further in the analysis by zooming on the 281 EU regions. In this regard, BIZMAP does not provide one unique solution but encourages the SME manager to explore and compare the different areas and criteria used to compute the scores. This user-friendly visualisation is particularly addressed to practitioners and can be understood without technical background.

The application builds on 7 different open access databases: Eurostat, OECD, World Bank, European Central Bank, European Investment Bank, European Commission, CEPII. After harmonization, we take advantage of machine learning methods to select the best predictors of bilateral flows of imports and FDI. In this model, the distance between the two countries as well as their gross domestic products are crucial for both types of flows. In addition, the legal framework regarding insolvency and the cost associated with border compliance and domestic transport play a significant role in bilateral flows of imports. Concerning FDI flows, taxes on goods and services as well as air freight are the most important factors.

By reducing informational uncertainty abroad, BIZMAP enhances traditional evaluation of commercial opportunities and enables SMEs to target some EU markets before launching
more accurate research. Hence, we enjoin the entrepreneur to use BIZMAP in complement with information coming from governmental agencies or sectoral market studies that could provide him more qualitative data.

This working paper aims at presenting in detail the methodology used in the application. It allows comments, suggestions and reactions to be collected from practitioners and researchers. In particular, one way to improve significantly the model would consist in using products classification, combined with countries, to model bilateral trade flows. Thus, the SME manager would be able to choose in the application not only the sector but also the product. Ultimately, BIZMAP is intended to be shared and used among SMEs which are seeking for new opportunities abroad.

Figure 1: Zoom on regional scores for a Portuguese SME in the construction sector
1. Introduction

In the post-1950 period, the global increase in the flows of trade, capital and information has helped push the world economy into a state of globalization, in which most of economies are highly interconnected (Masson, 2001). In this context, the firm-level internationalization refers to the expansion of international business operations such as exports, international partnerships or foreign direct investment (FDI). By fostering innovation and facilitating spillovers of technology, this participation in global markets may create opportunities to enhance productivity and can therefore be an important driver of employment growth (Wagner, 2012).

However, engaging in such activities can be expensive and usually only the most productive firms can afford to do so (Melitz, 2003; Helpman et al., 2004; Bernard et al., 2007). For instance, the entry into foreign markets implies transaction costs or fixed costs that can generate significant barriers (Eden & Miller, 2004). Given their small size, small and medium-sized enterprises (SMEs) suffer from typical obstacles which affect their ability to increase their activity abroad (Hollenstein, 2005; Paul et al., 2017). The latter can be classified as either internal, such as lack of internal resources, or external, such as uncertain institutional environments.

Hence, despite their importance in terms of activity and employment, SMEs only account for a small share of exports (OECD, 2015). In most OECD countries, for instance, SMEs represent more than 95% of all enterprises, about two-thirds of total employment and more than half of the value added of the business sector. Yet, their contribution to overall exports stands between 20% and 40% for most OECD economies (see figure 2).

Thus, although the fragmentation and specialization of global economic activity opens up a number of opportunities for SMEs, some obstacles to their exploitation remain. Among them, the lack of information is one of the most salient for example when it comes to selling goods and services on foreign markets (Lloyd-Reason et al., 2009). This patchy knowledge limits their ability to choose the geographical areas most suited to their business.

To overcome these difficulties, this paper combines many economic and financial data
in open access to determine a multidimensional indicator of the attractiveness of European territories. The latter makes it possible to evaluate, according to SMEs’ activity, which are the most promising markets based on a wide range of criteria. By reducing information uncertainty abroad, the BIZMAP application enables SMEs to improve their analysis of new markets through the use of an harmonized territorial attractiveness indicator.

To capture the protean nature of attractiveness at local level, the application builds on 7 different open access databases coming from Eurostat, the European Central Bank (ECB), the Organisation for Economic Cooperation and Development (OECD), the European Investment Bank (EIB), the European commission (AMECO), the World Bank and the Research and Expertise on the world economy (CEPII) which is a French institution specialized in international trade. Based on our expert judgment, we end up with an unified database encompassing more than 80 preselected variables for the 28 members of the European Union over the period 2015-2021.

Our approach relies on big data methods. First, we aggregate the time series and impute missing values with either random forest techniques or Kalman filters. Second, given the high dimensionality of our dataset, the most relevant variables are selected according to Lasso (Least Absolute Shrinkage Selection Operator) regressions applied to a gravity model of trade using either imports or FDI as dependent variable.

In the end, we obtain the contribution of each variable to exports or FDI in order to weight the variables we use to compute the indicators of geographical attractiveness and we propose a data visualization of our results inside our BIZMAP application. The principle is straightforward: after filling out all the necessary fields on the application (sector and home country), BIZMAP provides the SME with an European ranking based on an interactive visualization.

\footnote{Note that, for some variables, our dataset both incorporates the 3-year economic forecast of the European Commission and our own forecasts based on Kalman filter or random forest methods. See section 3 for more details.}
which indicates what are the most attractive European countries for its specific activity. Importantly, BIZMAP also enables SMEs to understand what are the main drivers of this score by presenting the contributions of the most important variables. Finally, BIZMAP offers the possibility to compare the 281 European regions with each other using the Eurostat NUTS 2 classification. Looking at countries heterogeneity, SMEs are therefore able to have a clearer picture of the most attractive European areas.

Our paper relates to the literature focusing on the firm decision to engage in international activity. While the traditional trade theories discuss the importance of differences in technology (David, 1817) and factor endowments (Heckscher & Ohlin, 1933) across economies to highlight comparative advantages, the New Trade Theory developed a model of monopolistic competition in which only the most productive firms internationalize their business (Melitz, 2003; Helpman et al., 2004). In contrast, we focus on the determinants of internationalization based on the economic potential of foreign markets. In particular, BIZMAP aims to reinforce the European economic integration which is likely to increase the growth potential of its members through higher regional trade (Vamvakidis, 1998).

The challenge of selecting the main drivers of the external performance among a wide range of possible variables was discussed in the economic growth literature under the so-called issue of "openendedness of theories" (Brock & Durlauf, 2001). In this case, one faces both the traditional problem of estimation uncertainty and the additional one of model uncertainty related to the choice of covariates. We tackle this issue by implementing Lasso methods, which provide a formal treatment of model uncertainty by considering all possible sets of variables.

---

3 The determinants of attractiveness are studied according to the Statistical classification of economic activities in the European Community (NACE Rev.2).

4 The current NUTS 2016 classification is valid from 1 January 2018 onwards and lists 104 regions at NUTS 1, 281 regions at NUTS 2 and 1348 regions at NUTS 3 level. The NUTS classification (Nomenclature of territorial units for statistics) is a hierarchical system for dividing up the economic territory of the EU for the purpose of socio-economic analyses of the regions. More information are available in the following page https://ec.europa.eu/eurostat/en/web/nuts/background.
The remainder of the paper is structured as follows. Section 2 presents the preselected potential drivers of internationalization. Section 3 deals with the empirical strategy. Section 4 discusses the results. Section 5 exhibits the BIZMAP application. Section 6 concludes.

2. Potential drivers of internationalization

The potential drivers of internationalization are numerous and incorporate among other aspects growth prospects, demography, education, quality of institutions, access to finance. Consequently, the latter can be searched in very wide areas of economics.

The first conceptual step consists in choosing the dependent variables which captures the economic potential of a foreign market. Here we focus on the balance of payments defining either flows of imports or FDI as measures of a country’s commercial dynamism. More precisely, we look at bilateral flows in order to evaluate trade potential of each European country with respect to a given country. Coming from the Eurostat database, bilateral flows of imports and FDI are timely and harmonized across European countries.

While constructing the set of potential drivers of internationalization, we considered the following themes: (i) economic prospects; (ii) infrastructure; (iii) institutional environment; (iv) financial conditions; and (v) demography and standard of living. The final dataset comprises 82 explanatory variables coming from 7 different open access databases over the period 2015-2021. The main categories of potential drivers of internationalization are discussed below (see Table 1 and Table 2 for names of variables, sources and the granularity of data available).

2.1. Trade related variables

This block of variables refers to various statistics that are informative for bilateral trade outcomes based on the detailed trade data of the French center of Research and Expertise on

Note that, for some variables, our dataset both incorporates the 3-year economic forecast of the European Commission and our own forecasts based on Kalman filter or random forest methods. See section ?? for more details.
the world economy (CEPII). Regarding the trade distance, we use the distance measures made available by the CEPII which hinge on city-level data to assess the geographic distribution of population (in 2004) inside each nation. The basic idea is to calculate distance between two countries based on bilateral distances between the biggest cities of those two countries, those inter-city distances being weighted by the share of the city in the overall country’s population. We also add two dummies: while the dummy trade contiguity takes the value 1 whether two countries are adjacent or 0 otherwise, the Common official language one takes the value 1 whether two countries share a common official language. In our case, we assume that a lower distance, a common language and a neighbouring country increase the probability of trading.

2.2. Economic prospects variables

To assess the economic potential of European countries, we rely on macroeconomic variables which describe the structure of the economy. Stemming from 4 different providers (the European commission, Eurostat, ECB, EIB), theses variables are available with different levels of granularity: country, macro-sector, NUTS 2 region, and NACE Rev. 2 activity.

As regards the database of the macro-economic database of the European Commission (AMECO), we select the gross value added at 2010 price, the harmonised consumer price index, the unemployment rate, the private final consumption expenditure and the gross fixed capital formation to account for productive capacity and growth prospects at the country-level. We also include the ECU-EUR exchange rates to take into account the effect of exchange rate volatility on countries attractiveness within the European Union.

Drawing on Eurostat, we complement these measures by the GDP and the unemployment growth rate at the regional level. We also use sector-specific variable such as labour costs or sentiment indicators. The latter are made up of five sectoral confidence indicators: industrial confidence indicator, services confidence indicator, consumer confidence indicator, construc-

---

tion confidence indicator and retail trade confidence indicator. At the NACE Rev 2 level, we include the amount of firm turnover, the wage adjusted labour productivity, the average personnel costs, the growth rate of employment, the gross operating surplus and the investment rate. In addition, we add the house price index as well as the amount of R&D expenditures at the national level.

Turning to financial variables we consider that foreign credit cycles may drive external demand. Thus, using the ECB database, the outstanding amounts of household and corporate credit are incorporated into the dataset. Finally, to capture the business cycle, we make use of the annual EIB Group Survey on Investment and Investment Finance (EIBIS). Encompassing all EU countries, this survey gathers qualitative and quantitative information on investment activities by small and medium-sized businesses and larger corporations, their financing requirements and the difficulties they face. It thus provides a wealth of unique firm-level information about investment decisions and investment finance choices. Restricting the survey to SMEs, we retain questions that focus on the expected investment, the share of companies that invest, the demand for products or services and the uncertainty about the future. Importantly, the EIBIS allows to gather the answers according to macro-sectors.

2.3. Institutional environment

The institutional environment of a given country plays a crucial role in attracting foreign firms. To proxy the quality of institutions and the rule of law which encourages international trade we make use of three different databases coming from the World Bank and the OECD. First we rely on the Worldwide Governance Indicators (WGI) which aggregate governance indicators for over 200 countries over the period 1996–2018, for six dimensions of governance: voice and Accountability, political Stability and absence of violence, government effectiveness, regulatory quality, rule of law, control of corruption. These aggregate indicators combine the views of a large number of enterprises, citizens and expert survey re-
spondents in industrial and developing countries. They are based on over 30 individual data sources produced by a variety of survey institutes, think tanks, non-governmental organizations, international organizations, and private sector firms.

Second, we choose the *Doing Business* indicators of the World Bank which provide objective measures of business regulations and their enforcement across 190 economies on the following topic: trading across borders, time to import, starting a business, resolving insolvency, regulatory quality, registering property, protecting minority investors and paying taxes.

Finally, we gather information about the tax environment of all European countries using the OECD tax database which provides comparative information on a range of tax statistics - tax revenues, personal income taxes, non-tax compulsory payments, corporate and capital income taxes and taxes on consumption - that are levied in the 35 OECD member countries.

2.4. *Infrastructure*

The quality of infrastructure is also of major importance in order to facilitate delivering freight from a given country to every country of the European Union. To proxy the beneficial effect of infrastructure, we add 8 more variables stemming from the World Bank, Eurostat and the EIB. While we select the variable *Getting electricity* of the *Doing Business* projet we also include transport network information available at the NUTS2 level in the Eurostat database such as the motorway network, the railway network, the air freight and the ocean freight. Then, exploiting the EIB’s survey on investment, we select questions dealing with energy costs, access to digital infrastructure and availability of adequate transport infrastructure.

2.5. *Financial conditions*

Since the onset of the crisis, financial variables have gained prominence in explaining the performance of both firms and countries. Based on Eurostat, the ECB database and the EIB’s survey on investment, we consider measures characterizing financing conditions including the debt of households, non-financial corporations and governments. In addition, we look at
the effect of financial stability measures such as non-performing loans, the country-level core tier one ratio of European banks or the financial stress indicator of the ECB. Besides we add measures of SMEs access to finance using answers of the EIB survey about the amount of credit obtained, the cost of the external finance obtained and even the collateral required.

2.6. Demography and standard of living

The trade attractiveness of a country is tightly connected with the demography and the standard of living. In particular, we include the Eurostat share of labour force with secondary and tertiary education, as well as answers about availability of staff with the right skills provided by the EIB survey to capture the skill endowment of the labour force. Besides, the total population or the level of inequality or poverty may also play an important role in determining the volume and the nature of the external demand. Finally, we complete the database with Eurostat information on environmental policies of the EU countries such as the share of renewable energy or the level of gas emissions.

3. Methodology

Building on this large amount of information, we take advantage of big data techniques to assess the relative importance of potential drivers of internationalization. To obtain an unified database, we first deal with the imputation of missing data for the 28 EU members. Depending on the nature of these data (partially or completely missing), two different algorithms are implemented. On the one hand, series where a year observation is missing are imputed using time series technique such as Kalman filtering (section 3.1.1). On the other hand, if the data are unavailable for a given geographical area (country or region) then multivariate imputation methods such as missForest are implemented to make use of the observed link between the

---

8 The Country Level Index of Financial Stress (CLIFS) includes six, mainly market-based, financial stress measures that capture three financial market segments: equity markets, bond markets and foreign exchange markets. In addition, when aggregating the sub-indices, the CLIFS takes the co-movement across market segments into account. See Duprey et al. (2017) for more details.
missing variable and the others in the areas where all data are available (section 3.1.2). From there, we model bilateral flows of imports and FDI through a gravity model of trade using a post-lasso OLS which consists in running an OLS on variables selected using a Lasso model.

### 3.1. Missing values imputation

As explained previously, our data are available at 4 different levels of aggregation: country, macro-sector, sector and region (see Figure 3). On each of the four levels, some data are missing. The first need is thus to impute these missing data. The Figure 4 show the average percentage of missing data for each year and each geographical area. For a given variable and a given geographical area (national or regional), the data can be either partially or totally missing. We address the first case using time serie techniques, while we rely on multivariate imputations for the second one. As a result, for each level, the missing data are imputed according to a specific algorithm (see Algorithm 1). The Figure 5 gives the count of the available observations and the imputed values according to the different techniques.

---

**Algorithm 1** Impute missing values for this level of granularity

```plaintext
for serie in this level do
    for area (nuts0 or nuts2) where some data is available do
        if enough observations (n ≥ 3) and non null variance then
            Impute using Kalman over the period 2015-2021
        else
            Impute using the mean value
        end if
    end for
    for area without available data do
        Impute using missForest with observations of the region with non missing values
    end for
end for
```

---

### 3.1.1. Kalman filtering

In this step, we complete missing values of time series taken individually. To do so, we model time series as state-space models based on a decomposition of the series into a number
of components (Harvey, 1990). The estimation of such state-space models is then done by a
Kalman filtering algorithm. The Figure 6 shows the output of such this algorithm for various
level of missingness.

More specifically, in our study, time series are modelled by a state-space model named
local linear model. This model is defined as follow:

\[ x_t = \mu_t + \epsilon_t, \quad \epsilon_t \sim N(0, \sigma_{\epsilon}^2) \] (1)

\[ \mu_{t+1} = \mu_t + \eta_t + \xi_t, \quad \xi_t \sim N(0, \sigma_{\xi}^2) \] (2)

\[ \eta_{t+1} = \eta_t + \zeta_t, \quad \zeta_t \sim N(0, \sigma_{\zeta}^2) \] (3)

where \( x_t \) are the observations defined as the sum of a time-varying slope \( \mu_t \) (unobserved)
and a noise \( \epsilon_t \) of variance \( \sigma_{\epsilon}^2 \). The time-varying slope is made of a random walk of variance \( \sigma_{\xi}^2 \)
which models the trend with and an additional random walk \( \eta_t \) of variance \( \sigma_{\zeta}^2 \) which models
the fact that the trend can vary over time. \( \epsilon_t \) is called the noise of the observations whereas \( \xi_t \)
and \( \zeta_t \) are called the noise of the model or the noise of the system. The Kalman algorithm is
used to estimate the variance of the three parameters \( \epsilon_t, \xi_t \) and \( \zeta_t \). Once these variances have
been estimated, the equation system of the state-space model enables us to complete missing
data in the time series.

3.1.2. missForest algorithm

Afterwards, we focus on the imputation of missing data where no observation is available
for a given geographical area (or a macro-sector/sector). For this purpose, we implement the
algorithm missForest (Stekhoven & Bühlmann, 2011), which is based on a random forest
predictor (Breiman, 2001). The benefits of using this classifier are numerous such as allowing
for interactive and non-linear effect, not relying on strong statistical hypothesis on the data,
providing prediction even though data are missing in its inputs (contrary to linear models for
instance), gracefully handling mixed data (such as categorical). Importantly, the absence of a-priori statistical hypotheses on the data is a desirable feature of missForest. Furthermore, one should note that missForest outperforms others standard approach with a decrease of imputation error of 50% in some cases (Stekhoven & Bühlmann [2011]).

This multivariate method consists in predicting the missing values using a random forest trained on the observed parts of the dataset. In other words, it makes use of all the other variables to predict the variable with missing values. At the first iteration of the imputation process, the missing data are imputed to their mean. The process stops as soon as the difference between the newly imputed data matrix and the previous one increases for the first time.

3.2. Trade flow modelling

3.2.1. A machine learning selection using Lasso

For now, we have preselected 82 variables based on our expert judgment. However, we still have an issue for the estimation of imports and FDI. Indeed, in situations where the dimensionality of the data may exceed the length of the sample size, overfitting concerns arise (Hawkins [2004]). In our database, variables are typically available at an annual frequency and available only for few years. In this case, least squares estimation cannot yield unique coefficient estimates and it is necessary to reduce the number of covariates included in the model. Consequently, before plugging all these variables in any model, we decide to go for a variable selection procedure.

Among the different existing methods, we implement Lasso (Least Absolute Shrinkage Selection Operator) regressions to set to zero covariates for which the absolute value of their estimates is lower than a level $\lambda$ (Tibshirani [1996]). The difference between a Least Squares regression and Lasso regression lies in the optimization problem solved. In fact, the Lasso regression adds a penalty term to the least squares term as follows:
\[
\min_{\theta} \sum_{i=1}^{n} \left( y_i - x_i \theta \right)^2 + \lambda \| \theta \|_1
\]

where \( y_i \) is the \( i^{th} \) observation of the independent variable, \( x_i \) denotes the covariates of the \( i^{th} \) observation, \( \theta \) corresponds to the estimates, \( \| \cdot \|_1 \) is the \( L_1 \) norm, and \( \lambda \) is the penalty parameter. The penalty parameter helps reducing the number of the covariates included in the model. The optimal \( \lambda \) is determined by cross-validation. The latter refers to a resampling technique which helps to find a parameter value that ensures a proper balance between bias and variance (or flexibility and interpretability).

The cross-validation used is the so-called K-fold cross-validation method that divides the dataset randomly into \( K \) different subsets. One subset is kept for validation while the model is estimated over the remaining \( K-1 \) subsets. This procedure is repeated for each subset and each \( \lambda \). The best penalty parameter value is the one yielding the lowest K-fold estimate. In our study, we chose a K-fold cross-validation with \( K=10 \) which is the default value for K-fold cross-validation.

Since the Lasso biases the coefficients towards zero, the estimates might not be consistent. This is even more true in presence of highly correlated covariates. Besides, Belloni et al. (2013) have shown that the post-Lasso OLS performs at least as well as the Lasso under mild additional assumptions. We therefore decide to use a two-step estimation procedure in which we regress our variables of interest on the subset of covariates chosen by the Lasso.

3.2.2. Gravity model equation

Once we have selected the most relevant variables according to the Lasso criteria, we then are able to incorporate those variables into a gravity model of trade. The gravity equation in international trade is one of the most robust empirical finding in economics (Chaney, 2018): bilateral trade between two countries is proportional to their respective sizes, measured by their GDP, and inversely proportional to the geographic distance between them. The traditional gravity model applied to bilateral trade flows is the following:
\[ X_{ij} = G \cdot \frac{Y_i^{\beta_1} Y_j^{\beta_2}}{D_{ij}^{\beta_3}} \]  

(5)

Where the trade flow \( X_{i,j} \) is explained by \( Y_i \) and \( Y_j \) that are the masses of the exporting and importing country (e.g. the GDP) and \( D_{ij} \) that is the distance between the countries. A logarithmic operator can be applied to form a log-linear model, which yields the following equation:

\[ \log X_{ij} = \beta_0 + \beta_1 \log Y_i + \beta_3 \log Y_j + \beta_4 \log D_{ij} + \epsilon_{ij} \]  

(6)

Additional bilateral variables such as contiguity (the fact that two countries share the same border), common language or regional trade agreement\(^9\) are often included in the equation. In the case of our dataset which includes more information, we estimate two different equations which are the following:

\[ \log M_{ij} = \beta_0 + \beta_1 \log Y_i + \beta_3 \log Y_j + \beta_4 \log D_{ij} + \beta_5 V_{ij} + \beta_6 Z_i + \epsilon_{ij} \]  

(7)

\[ \log FDI_{ij} = \beta_0 + \beta_1 \log Y_i + \beta_3 \log Y_j + \beta_4 \log D_{ij} + \beta_5 V_{ij} + \beta_6 Z_i + \epsilon_{ij} \]  

(8)

where the dependent variable is either \( M_{ij} \) and refers to the bilateral flow of imports of country \( i \) coming from country \( j \) or \( FDI_{ij} \) and represents the bilateral flow of FDI of country \( i \) coming from country \( j \). Besides, in both equations, \( Z_i \) is the matrix of potential drivers of attractiveness related to country \( i \) and \( V_{ij} \) is the matrix of bilateral variables between country \( i \) and \( j \) such as contiguity and common language.

\(^9\) Note that constant \( G \) becomes part of the \( \beta_0 \). Also for easier interpretation we decide not to invert the log of the distance, contrary to what would be implied by taking logarithm of Equation 5.

\(^{10}\) Since the countries under interest are in the Eurozone, this variable is de facto excluded.
3.2.3. Using different level of granularity

Finally, one last step of data processing is necessary before being able to run regressions. Indeed, since the explained variables are available at the country level only, the explanatory variable available at a more disaggregated level (i.e. macro-sector, sector or NUTS 2) must be aggregated to a country level. To this end, two ways of aggregating values are used: summation and product. The summation is used for the value that are absolute (i.e. not in percentage) whereas the product is used when the value is in percentage\[11\] In addition, for a prediction made at a more disaggregated level than the national level, only a subset of covariates are available for this level. Hence, for the variables not available at this level, we take the values of these variables at a more aggregated level. For instance, if the prediction is made at the sector level, the values we use for the variables not available at the sector level are the ones of the corresponding macro-sector. Similarly, if some variables are not available at the macro-sector level, then the values at the national level are retained.

4. Empirical results

4.1. Selected variable by Lasso

The gravity equations \[7\] and \[8\] are first estimated using Lasso regression. The selection process for the import equation retains 10 different variables, including the core variables of gravity models such as distance, the contiguity, the common language, the GDPs of the exporting and importing countries and their total population. With regards to the FDI equation, the Lasso regression selects 9 variables, and includes as well the core variables of the gravity model. In those two sets of selected variables, there are 5 common variables. Thus, even though some variables can explain both phenomena (imports and FDI flows), we still have some variables that do not overlap which means that some factors explaining each process are specific.

\[11\] Since the product is not weighted, note that we assume that each of the sector/macro-sector/region have the same weight.
4.2. Estimates from gravity models

The estimates of the variables selected in the Lasso are shown in Table 3 and Figure 7. In the latter, note that the explanatory variables are all standardized to allow direct comparison of the magnitudes of the effects. Regarding the bilateral FDI estimation in column (1) of Table 3, the three variables with the biggest effects are the GDPs of both countries, the logarithm of the distance and the importance of the air freight of the country attracting FDI. These effects are consistent and significant at the 1% level. An increase in the GDP of the investing country, as well as a decrease in the distance between the two countries lead to higher FDI. Also, better air transports in the receiving country are associated with higher FDI. Unsurprisingly, the coefficients of the GDP and the corporate credit of the country attracting FDI are both positive and significant. Indeed, these are overall demand factors that directly influence the investors decision to invest in a foreign country. In addition, from an investor perspective, higher level of taxation directly reduce financial profitability and thus has negative effect on investment volume. Hence, higher receiving country’s taxes on good and services have a negative effect on FDI. Conversely, sharing a common language has a positive impact on those inflows.

Regarding the estimations of imports flows, column (2) shows the same prominence of the core variables of the gravity model (GDPs of both countries and distance) that are all significant at the 1% level. Similarly, the coefficient related to air transports is still positive and significant. Yet, sharing a common language is no more significant while the contiguity variable turns out to increase the global volume of imports. Indeed, importing from a country might require less cultural proximity than investing in the long run through FDI. Instead, trading with a neighbouring country is of major importance even after controlling for the effect of the distance. Other variables related to institutions quality such as solvency rules or reduced time and costs associated with the logistical process of importing goods have also a positive and significant effect on these inflows. Furthermore, the share of corporate non-performing loans, which captures financial fragility, has a negative and significant effect on imports but its
magnitude is lower than the previous variables. Finally, turning to the demographic factors, the higher the population of the importing country, the higher the imports.

5. Implementation in BIZMAP

5.1. Software and hardware used

BIZMAP is a web application built within the shiny framework in R. Regarding the user interface, the core package shiny combines shinydashboard and shinydashboardPlus to enhance the user experience. In order to guide the user, a tutorial has been created with the package rintrojs. The latter is available on the menu help of the application. Furthermore, some custom CSS and JS scripts have been developed to enhance style and dynamics of the application. Turning to the web infrastructure, the application has been deployed on an Amazone EC2 instance with the following configuration: variable ECU, 2 vCPU, 2.3 GHz, Intel Broadwell E5-2686v4, 8 Go memory, EBS only. A shiny server has also been installed and configured on the AWS instance to receive the application locally developed.

5.2. Operating instructions

When an user opens the app, he has access to a left menu where he is asked to fill in some information. First, the user has to determine whether he wants to export or make a Foreign Direct Investment for his business as described in Figure 8. Depending on whether the user is interested in current indicators or predictions, he has to choose the period he is interested in within the 2015-2021 period (see Figure 9). Then, the user has to select the country where his company is located as shown in Figure 10. In fact, our indicators rely on geographical distances between this localization and all other EU countries. Finally, the user has to fill in the macro-sector and the sector of his company to obtain results tailored to his business (see Figure 11). All in all, it is possible to choose among a total of 21 macro-sectors and 88 sectors.

Once the left menu is completed, models are running and indicators for each countries and each regions are computed. The application is buffering layers based on the value of the
indicators. Values are scaled between 0 and 100 (see Figure 12) and the higher the value, the better the user has interest to export or make a Foreign Direct Investment. Graphically, the most attractive European territories are represented with the warmest colors.

For example, consider the case of a Portuguese SME specialized in the retail of Portuguese wine that intend to export its production in Europe. The firm wants to know where are the best opportunities in Europe for its products so it fills in the information needed in the left menu presented in Figure 13. From there, the application provides the user with a ranking of the Top 10 best countries to export (see Figure 14). The application also enables to have a global view of the scores of all the countries of the European Union (see Figure 15). In our example, the SME should export to France (100), Spain (97.9) and United Kingdom (89.6). Importantly, it is possible to display the score of any country by hovering the mouse over it. In addition, by clicking on a country, the user has access to an in-depth analysis that explains the scores obtained through the visualization of the contribution of each theme to the score (see Figure 16). In the case of our Portuguese entrepreneur, the SME has interest in exporting to France mainly because of a better demography, a higher standard of living, a well developed infrastructure and a strong institutional environment.

The user has also access to even deeper analysis with the Analytics menu on top of the app. First, the number of best countries can be selected from 1 to 10 in order to display the ranking of the top countries (see Figure 17). Second, each country score can be broken down into our five different themes (see Figure 18) or the top 8 most impacting variables (see Figure 19). Figure 17 displays the same information available in the table exhibiting the ranking. However, a dashed line is added to represent the average score of all EU countries in order to enable a cross-country comparison between the different scores. Figure 18 displays the same information available on the map by clicking on a country but here, the information is displayed for the top countries all together making any comparison easier. A point for each theme is added to represent the mean contribution of a theme across all countries in order to have a better idea of the significance of the difference between the values. Besides, Figure 19
presents some new analytics. For each country, the contribution of the 8 most impacting variables is presented with the mean contribution of each variables across all countries. Again, this reference point enables to have a better evaluation of any value.

Finally, the application allows to obtain all the previous result at the regional level. The user only has to go back to the map and scroll up to zoom. Then, BIZMAP updates all the predictions to compute the score of territorial attractiveness at the regional level. Returning to our Portuguese example, Figure 20 presents the new ranking at the regional level, while Figure 21 exhibits the scores of all the EU regions. Once again, the user can go to the Analytics part to explore the results at the regional level.

6. Conclusion

In this paper, we have built a web application (BIZMAP) that enables SMEs to improve their analysis of foreign markets through the use of an harmonized territorial attractiveness indicator. Many challenges arise from the construction of such application. As a matter of fact, starting from the collection and manipulation of big data provided by 7 different open access databases, we deal with missing values and aggregation issues. From there we start using machine learning methods such as random forest and Kalman filtering. Then, we used a two-steps estimation procedure by combining Lasso regression to select the most relevant variables and a gravity model of trade to determine the most attractive region of Europe. Last but not least, we end up with a synthetic indicator easily readable by SMEs to help them in their decision-making process.

Our datascience pipeline enables us to build a flexible application that covers all the 28 members of the European Union at both a national and regional level, thus providing an indicator about the territorial attractiveness concerning 21 macro-sectors and 88 sectors from 2015 to 2021. Hence, by reducing information uncertainty abroad, BIZMAP is likely to improve the SMEs’ analysis of new markets through the visualization of harmonized territorial attractiveness indicators.
This project can still be improved in many ways. More data available at a NUTS 2 level concerning sectors will reduce the number of artificial completions that we made, thus making our models more reliable. Moreover, data more related to the core business of the SMEs would be more useful for them. So, it might be interesting to make an analysis at a deeper sectoral level. Another way to improve the application is to explore the residuals of the gravity models in order to understand which factors are missing. Finally, we plan to extend the application to a wider range of countries, for instance the members of the OECD.

References


Heckscher, E., & Ohlin, B. (1933). Factor-endowment and factor proportion theory.


Lloyd-Reason, L., Ibeh, K., & Deprey, B. (2009). Top barriers and drivers to sme internationalisation,


Figure 2: Economic importance of SMEs as compared to their contribution in global trade

SME export activity, value added and employment shares, as a percentage, 2013

Source: OECD Structural and Demographic Business Statistics and Trade by Enterprise Characteristics databases.

Figure 3: The various levels of granularity

lower granularity

National (NUTS0) level

Macrosectors level

Region (NUTS2) level

higher granularity

Sectors level
<table>
<thead>
<tr>
<th>Variable</th>
<th>Units</th>
<th>Database</th>
<th>Granularity</th>
<th>Theme</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bilateral flows of imports</td>
<td>Million €</td>
<td>Eurostat</td>
<td>Country</td>
<td>Trade</td>
</tr>
<tr>
<td>Bilateral flows of FDI</td>
<td>Million €</td>
<td>Eurostat</td>
<td>Country</td>
<td>Trade</td>
</tr>
<tr>
<td>Bilateral trade distance (weighted)</td>
<td>Km</td>
<td>CEPII</td>
<td>Country</td>
<td>Trade</td>
</tr>
<tr>
<td>Trade contiguity</td>
<td>-</td>
<td>CEPII</td>
<td>Country</td>
<td>Trade</td>
</tr>
<tr>
<td>Common official language</td>
<td>-</td>
<td>CEPII</td>
<td>Country</td>
<td>Trade</td>
</tr>
<tr>
<td>Total population</td>
<td>Number</td>
<td>Eurostat</td>
<td>NUTS 2 region</td>
<td>Demography and standard of living</td>
</tr>
<tr>
<td>Young population</td>
<td>Number</td>
<td>Eurostat</td>
<td>NUTS 2 region</td>
<td>Demography and standard of living</td>
</tr>
<tr>
<td>Household income</td>
<td>Million €</td>
<td>Eurostat</td>
<td>NUTS 2 region</td>
<td>Demography and standard of living</td>
</tr>
<tr>
<td>Poverty rate</td>
<td>%</td>
<td>Eurostat</td>
<td>NUTS 2 region</td>
<td>Demography and standard of living</td>
</tr>
<tr>
<td>Share of renewable energy</td>
<td>%</td>
<td>Eurostat</td>
<td>Country</td>
<td>Demography and standard of living</td>
</tr>
<tr>
<td>Income share of the bottom 40%</td>
<td>%</td>
<td>Eurostat</td>
<td>Country</td>
<td>Demography and standard of living</td>
</tr>
<tr>
<td>Greenhouse gas emission</td>
<td>Tonnes per capita</td>
<td>Eurostat</td>
<td>Country</td>
<td>Demography and standard of living</td>
</tr>
<tr>
<td>Women in senior management position</td>
<td>%</td>
<td>Eurostat</td>
<td>Country</td>
<td>Demography and standard of living</td>
</tr>
<tr>
<td>High educational level</td>
<td>% of positions</td>
<td>Eurostat</td>
<td>Country</td>
<td>Demography and standard of living</td>
</tr>
<tr>
<td>Availability of staff with the right skills : major obstacle</td>
<td>%</td>
<td>European Investment Bank</td>
<td>Macro-sector</td>
<td>Demography and standard of living</td>
</tr>
<tr>
<td>Gross Value Added at 2010 prices</td>
<td>Billion €</td>
<td>AMECO</td>
<td>Country</td>
<td>Economic prospects</td>
</tr>
<tr>
<td>Harmonised consumer price index</td>
<td>Index</td>
<td>AMECO</td>
<td>Country</td>
<td>Economic prospects</td>
</tr>
<tr>
<td>Unemployment rate</td>
<td>%</td>
<td>AMECO</td>
<td>Country</td>
<td>Economic prospects</td>
</tr>
<tr>
<td>Private final consumption expenditure</td>
<td>Billion €</td>
<td>AMECO</td>
<td>Country</td>
<td>Economic prospects</td>
</tr>
<tr>
<td>Gross fixed capital formation</td>
<td>Billion €</td>
<td>AMECO</td>
<td>Country</td>
<td>Economic prospects</td>
</tr>
<tr>
<td>ECU-EUR exchange rates</td>
<td>Number</td>
<td>AMECO</td>
<td>Country</td>
<td>Economic prospects</td>
</tr>
<tr>
<td>GDP</td>
<td>Billion €</td>
<td>Eurostat</td>
<td>NUTS 2 region</td>
<td>Economic prospects</td>
</tr>
<tr>
<td>Sentiment indicators</td>
<td>Index</td>
<td>Eurostat</td>
<td>Macro-sector</td>
<td>Economic prospects</td>
</tr>
<tr>
<td>Consumer Sentiment indicators</td>
<td>Index</td>
<td>Eurostat</td>
<td>Country</td>
<td>Economic prospects</td>
</tr>
<tr>
<td>Unemployment rate</td>
<td>%</td>
<td>Eurostat</td>
<td>NUTS 2 region</td>
<td>Economic prospects</td>
</tr>
<tr>
<td>Household credit</td>
<td>Billion €</td>
<td>European Central Bank</td>
<td>Country</td>
<td>Economic prospects</td>
</tr>
<tr>
<td>NFC credit</td>
<td>Billion €</td>
<td>European Central Bank</td>
<td>Country</td>
<td>Economic prospects</td>
</tr>
<tr>
<td>Labor costs index</td>
<td>Index</td>
<td>Eurostat</td>
<td>NACE Rev. 2 activity (2 digit)</td>
<td>Economic prospects</td>
</tr>
<tr>
<td>House price index</td>
<td>Index</td>
<td>Eurostat</td>
<td>Country</td>
<td>Economic prospects</td>
</tr>
<tr>
<td>R&amp;D expenditures</td>
<td>Billion €</td>
<td>Eurostat</td>
<td>Country</td>
<td>Economic prospects</td>
</tr>
<tr>
<td>Expected investment : increase</td>
<td>%</td>
<td>European Investment Bank</td>
<td>Macro-sector</td>
<td>Economic prospects</td>
</tr>
<tr>
<td>Share of companies that invest : increase</td>
<td>%</td>
<td>European Investment Bank</td>
<td>Macro-sector</td>
<td>Economic prospects</td>
</tr>
<tr>
<td>Demand for product or service : major obstacle</td>
<td>%</td>
<td>European Investment Bank</td>
<td>Macro-sector</td>
<td>Economic prospects</td>
</tr>
<tr>
<td>Uncertainty about the future: major obstacle</td>
<td>%</td>
<td>European Investment Bank</td>
<td>Macro-sector</td>
<td>Economic prospects</td>
</tr>
<tr>
<td>Turnover</td>
<td>Billion €</td>
<td>Eurostat</td>
<td>NACE Rev. 2 activity (2 digit)</td>
<td>Economic prospects</td>
</tr>
<tr>
<td>Wage adjusted labour productivity</td>
<td>%</td>
<td>Eurostat</td>
<td>NACE Rev. 2 activity (2 digit)</td>
<td>Economic prospects</td>
</tr>
<tr>
<td>Average personnel costs (personnel costs per employee)</td>
<td>Thousand €</td>
<td>Eurostat</td>
<td>NACE Rev. 2 activity (2 digit)</td>
<td>Economic prospects</td>
</tr>
<tr>
<td>Growth rate of employment</td>
<td>%</td>
<td>Eurostat</td>
<td>NACE Rev. 2 activity (2 digit)</td>
<td>Economic prospects</td>
</tr>
<tr>
<td>Gross operating rate (gross operating surplus/turnover)</td>
<td>%</td>
<td>Eurostat</td>
<td>NACE Rev. 2 activity (2 digit)</td>
<td>Economic prospects</td>
</tr>
<tr>
<td>Investment rate (investment/value added at factors cost)</td>
<td>%</td>
<td>Eurostat</td>
<td>NACE Rev. 2 activity (2 digit)</td>
<td>Economic prospects</td>
</tr>
</tbody>
</table>

Notes: Continues on the next page.
Table 2: List of variables (continued)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Units</th>
<th>Database</th>
<th>Granularity</th>
<th>Theme</th>
</tr>
</thead>
<tbody>
<tr>
<td>Household debt (% of GDP)</td>
<td>%</td>
<td>Eurostat</td>
<td>Country</td>
<td>Financial conditions</td>
</tr>
<tr>
<td>NFC debt (% of GDP)</td>
<td>%</td>
<td>Eurostat</td>
<td>Country</td>
<td>Financial conditions</td>
</tr>
<tr>
<td>Public debt (% of GDP)</td>
<td>%</td>
<td>Eurostat</td>
<td>Country</td>
<td>Financial conditions</td>
</tr>
<tr>
<td>Core tier one ratio</td>
<td></td>
<td>European Central Bank</td>
<td>Country</td>
<td>Financial conditions</td>
</tr>
<tr>
<td>Financial stress indicator</td>
<td>Index</td>
<td>European Central Bank</td>
<td>Country</td>
<td>Financial conditions</td>
</tr>
<tr>
<td>Household non-performing loans</td>
<td>%</td>
<td>European Central Bank</td>
<td>Country</td>
<td>Financial conditions</td>
</tr>
<tr>
<td>Corporate non-performing loans</td>
<td>%</td>
<td>European Central Bank</td>
<td>Country</td>
<td>Financial conditions</td>
</tr>
<tr>
<td>Availability of finance: major obstacle</td>
<td>%</td>
<td>European Investment Bank</td>
<td>Macro-sector</td>
<td>Financial conditions</td>
</tr>
<tr>
<td>The amount of credit obtained: dissatisfied</td>
<td>%</td>
<td>European Investment Bank</td>
<td>Macro-sector</td>
<td>Financial conditions</td>
</tr>
<tr>
<td>The cost of the external finance you obtained: dissatisfied</td>
<td>%</td>
<td>European Investment Bank</td>
<td>Macro-sector</td>
<td>Financial conditions</td>
</tr>
<tr>
<td>The collateral required: dissatisfied</td>
<td>%</td>
<td>European Investment Bank</td>
<td>Macro-sector</td>
<td>Financial conditions</td>
</tr>
<tr>
<td>Getting electricity</td>
<td>Index</td>
<td>World Bank</td>
<td>Country</td>
<td>Infrastructure</td>
</tr>
<tr>
<td>Motorway network</td>
<td>Km</td>
<td>Eurostat</td>
<td>Country</td>
<td>Infrastructure</td>
</tr>
<tr>
<td>Railway network</td>
<td>Km</td>
<td>Eurostat</td>
<td>Country</td>
<td>Infrastructure</td>
</tr>
<tr>
<td>Air freight</td>
<td>Thousand tonnes</td>
<td>Eurostat</td>
<td>Country</td>
<td>Infrastructure</td>
</tr>
<tr>
<td>Ocean freight</td>
<td>Thousand tonnes</td>
<td>Eurostat</td>
<td>Country</td>
<td>Infrastructure</td>
</tr>
<tr>
<td>Energy costs</td>
<td>Index</td>
<td>European Investment Bank</td>
<td>Macro-sector</td>
<td>Infrastructure</td>
</tr>
<tr>
<td>Access to digital infrastructure: major obstacle</td>
<td>Index</td>
<td>European Investment Bank</td>
<td>Macro-sector</td>
<td>Infrastructure</td>
</tr>
<tr>
<td>Availability of adequate transport infrastructure: major obstacle</td>
<td>%</td>
<td>European Investment Bank</td>
<td>Macro-sector</td>
<td>Infrastructure</td>
</tr>
<tr>
<td>Voice and Accountability</td>
<td>Index</td>
<td>World Bank</td>
<td>Country</td>
<td>Institutional environment</td>
</tr>
<tr>
<td>Trading across borders</td>
<td>Index</td>
<td>World Bank</td>
<td>Country</td>
<td>Institutional environment</td>
</tr>
<tr>
<td>Time to import</td>
<td>Index</td>
<td>World Bank</td>
<td>Country</td>
<td>Institutional environment</td>
</tr>
<tr>
<td>Starting a business</td>
<td>Index</td>
<td>World Bank</td>
<td>Country</td>
<td>Institutional environment</td>
</tr>
<tr>
<td>Rule of Law</td>
<td>Index</td>
<td>World Bank</td>
<td>Country</td>
<td>Institutional environment</td>
</tr>
<tr>
<td>Resolving insolvency</td>
<td>Index</td>
<td>World Bank</td>
<td>Country</td>
<td>Institutional environment</td>
</tr>
<tr>
<td>Regulatory Quality</td>
<td>Index</td>
<td>World Bank</td>
<td>Country</td>
<td>Institutional environment</td>
</tr>
<tr>
<td>Registering property</td>
<td>Index</td>
<td>World Bank</td>
<td>Country</td>
<td>Institutional environment</td>
</tr>
<tr>
<td>Protecting minority investors</td>
<td>Index</td>
<td>World Bank</td>
<td>Country</td>
<td>Institutional environment</td>
</tr>
<tr>
<td>Political Stability</td>
<td>Index</td>
<td>World Bank</td>
<td>Country</td>
<td>Institutional environment</td>
</tr>
<tr>
<td>Paying taxes</td>
<td>Index</td>
<td>World Bank</td>
<td>Country</td>
<td>Institutional environment</td>
</tr>
<tr>
<td>Government Effectiveness</td>
<td>Index</td>
<td>World Bank</td>
<td>Country</td>
<td>Institutional environment</td>
</tr>
<tr>
<td>Getting credit</td>
<td>Index</td>
<td>World Bank</td>
<td>Country</td>
<td>Institutional environment</td>
</tr>
<tr>
<td>Enforcing contracts</td>
<td>Index</td>
<td>World Bank</td>
<td>Country</td>
<td>Institutional environment</td>
</tr>
<tr>
<td>Control of Corruption</td>
<td>Index</td>
<td>World Bank</td>
<td>Country</td>
<td>Institutional environment</td>
</tr>
<tr>
<td>Social security contributions (% of GDP)</td>
<td>%</td>
<td>OECD tax database</td>
<td>Country</td>
<td>Institutional environment</td>
</tr>
<tr>
<td>Tax on corporate profit (% of GDP)</td>
<td>%</td>
<td>OECD tax database</td>
<td>Country</td>
<td>Institutional environment</td>
</tr>
<tr>
<td>Tax on payroll (% of GDP)</td>
<td>%</td>
<td>OECD tax database</td>
<td>Country</td>
<td>Institutional environment</td>
</tr>
<tr>
<td>Tax on goods and services (% of GDP)</td>
<td>%</td>
<td>OECD tax database</td>
<td>Country</td>
<td>Institutional environment</td>
</tr>
<tr>
<td>Labour market regulations: major obstacle</td>
<td>%</td>
<td>European Investment Bank</td>
<td>Macro-sector</td>
<td>Institutional environment</td>
</tr>
<tr>
<td>Business regulations: major obstacle</td>
<td>%</td>
<td>European Investment Bank</td>
<td>Macro-sector</td>
<td>Institutional environment</td>
</tr>
</tbody>
</table>
Figure 4: Percentage of missing value as function of year and country
Figure 5: Available and imputed data
Figure 6: Kalman filtering examples

Explanation: The graph show the output of the Kalman filtering for different levels of missing values (from 9.1% to 57.1%). The solid black lines show the original series and the grey dotted line the Kalman filtering extrapolations.
Table 3: Estimations of the gravity model of imports and FDI

<table>
<thead>
<tr>
<th>Model</th>
<th>FDI (1)</th>
<th>Imports (2)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Trade</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Contiguity</td>
<td>0.134</td>
<td>0.132***</td>
</tr>
<tr>
<td>Log weighted distance</td>
<td>-0.501***</td>
<td>-0.824***</td>
</tr>
<tr>
<td>Common official language</td>
<td>0.151**</td>
<td></td>
</tr>
<tr>
<td><strong>Economic prospects</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Origin: GDP</td>
<td>1.048***</td>
<td>1.183***</td>
</tr>
<tr>
<td>Destination: GDP</td>
<td>0.347*</td>
<td>0.377***</td>
</tr>
<tr>
<td>Consumer sentiment indicators</td>
<td>0.123</td>
<td></td>
</tr>
<tr>
<td><strong>Institutional environment</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Resolving insolvency</td>
<td>0.099***</td>
<td></td>
</tr>
<tr>
<td>Taxes on good and services</td>
<td>-0.224**</td>
<td></td>
</tr>
<tr>
<td>Trading across borders</td>
<td>0.133***</td>
<td></td>
</tr>
<tr>
<td><strong>Infrastructure</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Air Freight</td>
<td>0.562***</td>
<td>0.238***</td>
</tr>
<tr>
<td>Railway network</td>
<td>0.038</td>
<td></td>
</tr>
<tr>
<td><strong>Financial conditions</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corporate NPL</td>
<td>-0.054**</td>
<td></td>
</tr>
<tr>
<td>Corporate credit</td>
<td>0.370***</td>
<td></td>
</tr>
<tr>
<td><strong>Demography and standard of living</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Destination: total population</td>
<td>0.443***</td>
<td></td>
</tr>
</tbody>
</table>

Observations 896 1,486  
Countries 28 28  
R² 0.441 0.900  
Adjusted R² 0.435 0.899  

Notes: The table shows the results of equation 7 and 8. All variable definitions are presented in Table 1 and 2. *, ** and *** indicate significance levels at 10%, 5% and 1% respectively.
Figure 7: Normalized gravity model estimations

Notes: The figure shows the results of equation 7 and 8. *, ** and *** indicate significance levels at 10%, 5% and 1% respectively. All explanatory variables are normalized using standardization (i.e. variables are centered and reduced) to allow direct comparison of the effects.
Figure 8: Choice between exportation and Foreign Direct Investment
Figure 9: Choice of the period
Figure 10: Choice of the country
Figure 11: Choice of the macro-sector and sector

Figure 12: Scale of the indicator displayed on the application
Figure 13: Informations filled by a Portuguese specialized in the retail of Portuguese wine willing to export

![Bizmap interface](image)

Figure 14: Top 10 best countries to export for the Portuguese company

<table>
<thead>
<tr>
<th>Rank</th>
<th>Country</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>France</td>
<td>100.0</td>
</tr>
<tr>
<td>2</td>
<td>Spain</td>
<td>97.9</td>
</tr>
<tr>
<td>3</td>
<td>United Kingdom</td>
<td>90.6</td>
</tr>
<tr>
<td>4</td>
<td>Germany</td>
<td>88.3</td>
</tr>
<tr>
<td>5</td>
<td>Italy</td>
<td>85.5</td>
</tr>
<tr>
<td>6</td>
<td>Belgium</td>
<td>77.0</td>
</tr>
<tr>
<td>7</td>
<td>Netherlands</td>
<td>73.6</td>
</tr>
<tr>
<td>8</td>
<td>Poland</td>
<td>66.4</td>
</tr>
<tr>
<td>9</td>
<td>Sweden</td>
<td>64.3</td>
</tr>
<tr>
<td>10</td>
<td>Denmark</td>
<td>64.2</td>
</tr>
</tbody>
</table>
Figure 15: Scores of all the countries of the European Union
Figure 16: Contribution of each theme to the score

France is ranked 1 with a score of 100
Figure 17: Ranking of the top countries
Figure 18: Contribution of each theme to the score

Score by theme

- France
- Spain
- United Kingdom

Institutional environment
Infrastructure
Financial conditions
Economic prospects
Demography and standard of liv

Score

Average value per theme

Figure 19: Contribution of each 8 most impacting variables

Score by variable

- France
- Spain
- United Kingdom

- Germany
- Italy
- Belgium

Score

Average value per variable
Figure 20: Top 20 best regions to export for the Portuguese company

<table>
<thead>
<tr>
<th>Rank</th>
<th>Region</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Andalucía</td>
<td>100.0</td>
</tr>
<tr>
<td>2</td>
<td>Ile-de-France</td>
<td>95.4</td>
</tr>
<tr>
<td>3</td>
<td>Lombardia</td>
<td>84.2</td>
</tr>
<tr>
<td>4</td>
<td>Zuid-Holland</td>
<td>82.6</td>
</tr>
<tr>
<td>5</td>
<td>Cataluña</td>
<td>81.7</td>
</tr>
<tr>
<td>6</td>
<td>Comunidad de Madrid</td>
<td>81.6</td>
</tr>
<tr>
<td>7</td>
<td>Rhône-Alpes</td>
<td>81.2</td>
</tr>
<tr>
<td>8</td>
<td>Comunidad Valenciana</td>
<td>74.0</td>
</tr>
<tr>
<td>9</td>
<td>Castilla y León</td>
<td>73.9</td>
</tr>
<tr>
<td>10</td>
<td>Provence-Alpes-Côte d'Azur</td>
<td>73.1</td>
</tr>
<tr>
<td>11</td>
<td>Galicia</td>
<td>72.9</td>
</tr>
<tr>
<td>12</td>
<td>Extremadura</td>
<td>72.7</td>
</tr>
<tr>
<td>13</td>
<td>Aquitaine</td>
<td>72.0</td>
</tr>
<tr>
<td>14</td>
<td>Castilla-La Mancha</td>
<td>71.0</td>
</tr>
<tr>
<td>15</td>
<td>Prov. Antwerpen</td>
<td>70.9</td>
</tr>
<tr>
<td>16</td>
<td>Pays de la Loire</td>
<td>70.0</td>
</tr>
<tr>
<td>17</td>
<td>Nord-Pas de Calais</td>
<td>69.2</td>
</tr>
<tr>
<td>18</td>
<td>Centre - Val de Loire</td>
<td>66.8</td>
</tr>
<tr>
<td>19</td>
<td>Midi-Pyrénées</td>
<td>66.7</td>
</tr>
<tr>
<td>20</td>
<td>Languedoc-Roussillon</td>
<td>65.7</td>
</tr>
</tbody>
</table>
Figure 21: Scores of all the regions of the European Union
Fostering European SMEs’ internationalization using Big Data: the BIZMAP application

IFC workshop on Data Science in central banking

Chloe Brochet Lostie de Kerhor
Yasmine Houri
Jean-Noël Kien
Etienne Kintzler
Lou Richardet

ILIAD-DDSA-DGSI
BANQUE DE FRANCE

20 OCTOBER 2021

The views expressed in the presentation are the sole responsibility of the authors and do not necessarily represent those of the Banque de France or the Eurosystem. All remaining errors are our own responsibility.
1. Introduction

2. Overview

3. Methodology

4. Results

5. Perspectives
1. INTRODUCTION

**Birth:** 2019 EU Datathon of the European Commission - 3rd prize

**Ambition:** to help companies to export (in particular SMEs)
- Economic weight of SMEs (95% of companies, 50% of employment)
- Obstacle to export: lack of resources, especially information (30% of exports)

**Solution:** identify the attractiveness for exports of European territories

**Use cases:**
- Visualize a model
- Help a French company to define an export strategy
- Demonstrate France's attractiveness abroad
- Provide quantitative data to experts
- Advise public authorities
1. Introduction

2. Overview

3. Methodology

4. Results

5. Perspectives
2. OVERVIEW

DATA
Multidimensional harmonised database

MODEL
Hybrid predictive model of international economics and machine learning

APPLICATION
Index of attractiveness of European territories

80 variables divided into 6 themes
- Institutional environment
- Economic perspectives
- Infrastructure
- Standard of living
- Financial conditions
- Geographical and cultural distance

3 dimensions
- Time (years)
- Spatial (countries, regions)
- Sectoral

Imputation of missing values
missForest + Kallman filters

Model
- Gravity model
- Augmented with ML: lasso
- Predictive
- Calibrated on export flows
- Provides attractiveness scores

Index of attractiveness
- Multidimensional, sectoral thanks to the database
- Possibility to look at future attractiveness through the model

Visualisations of the indicator
- Dynamics
- Ranking of territories
- Map with zoom on regions
- Contribution of the variables to the indicator

Need for interpretability and pedagogy

Allows a company to determine a custom export strategy
1. Introduction

2. Overview

3. Methodology

4. Results

5. Perspectives
### 3. METHODOLOGY

**IMPUTATION OF MISSING VALUES**

#### 1. Completion **between countries** with missForest*

- **MissForest** = nonparametric missing value imputation using random forest

#### 2. (For each country) completion **through time** with Kalman filters

---

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>AT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ES</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>AT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DE</td>
<td></td>
<td></td>
<td></td>
<td>missForest</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ES</td>
<td></td>
<td></td>
<td></td>
<td>missForest</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>missForest</td>
<td></td>
</tr>
<tr>
<td>NL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>AT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Kalman filtering</td>
<td>Kalman filtering</td>
</tr>
<tr>
<td>BE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Kalman filtering</td>
<td>Kalman filtering</td>
</tr>
<tr>
<td>DE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Kalman filtering</td>
<td>Kalman filtering</td>
</tr>
<tr>
<td>ES</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Kalman filtering</td>
<td>Kalman filtering</td>
</tr>
<tr>
<td>FR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Kalman filtering</td>
<td>Kalman filtering</td>
</tr>
<tr>
<td>IT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Kalman filtering</td>
<td>Kalman filtering</td>
</tr>
<tr>
<td>NL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Kalman filtering</td>
<td>Kalman filtering</td>
</tr>
</tbody>
</table>
3. METHODOLOGY
ESTIMATED EQUATION

- Standard gravity model augmented with economic variables

\[ Y_{ij}^k = \beta_0 + \sum G_{ij} + GDP_i + GDP_j + \sum X_j^k \]

- \( Y_{ij}^k \): exports from country \( i \) to country \( j \) for sector \( k \)
- \( G_{ij} \) & \( GDP_i \) or \( j \): variables of the standard gravity model
  - geographical, bilateral and cultural variables
  - economic mass of countries
- \( X_j^k \): Economic variables for country \( j \) and sector \( k \) (if available)
  - Institutional environment
  - Economic outlook
  - Infrastructure
  - Life standards
  - Financial conditions
3. METHODOLOGY
ESTIMATION OLS POST-LASSO BY SECTORS

Gravity model variables (G) + 80 economic variables (X) selected by experts

1. Variable selection with LASSO

2. OLS regression on all severity variables + selected economic variables

⇒ Gravity variables are always selected
⇒ Different economics variables are selected for different sectors

Reference: Least squares after model selection in high-dimensional sparse model (BELLONI and CHERNOZHUKOV, 2013)
https://www.jstor.org/stable/23525734?seq=1
1. Introduction

2. Overview

3. Methodology

4. Results

5. Perspectives
4. RESULTS
COEFFICIENTS PER SECTOR

- The manufacturing sector is the biggest macrosector, it includes many subsectors (food, beverage, wood, iron...): many variables are selected
- The arts and sciences are smaller and more specific sectors: fewer variables are selected
- The severity variables are robust
4. RESULTS
MACHINE LEARNING IMPROVES PERFORMANCE

- Performance: random forest (RF) >> hybrid model > standard gravity model
- Machine learning (RF) is more accurate but more difficult to interpret:
  - Trade-off between performance and interpretability
  - Specific methods can be used: feature importance, partial dependence plot, shapley value
- Other models can be tested: Gradient boosting (XGBoost, Catboost, etc.)
1. Introduction

2. Overview

3. Methodology

4. Results

5. Perspectives
Recently completed project
- Automate the update of data

Current work
- Take into account the effects of the health crisis (adapt the model)
- Testing other ML models
- Facilitating the interpretation of attractiveness scores
- Adding information and visualisations to the application
- Exchange with companies and international trade experts to improve the application

Possible developments
- Looking at the attractiveness of French regions to each other
- Create indicators of market potential / business survival