Statistic regression and open data approach for identifying economic indicators that influence e-commerce

Apollinaire Barme, Simon Tamayo, Arthur Gaudron

Statistic regression and open data approach for identifying economic indicators that influence e-commerce

Apollinaire Barme, Simon Tamayo, Arthur Gaudron

***Abstract*—** This paper presents a statistical approach to identify explanatory variables linearly related to e-commerce sales. The proposed methodology allows specifying a regression model in order to quantify the relevance between openly available data (economic and demographic) and national e-commerce sales. The proposed methodology consists in collecting data, preselecting input variables, performing regressions for choosing variables and models, testing and validating. The usefulness of the proposed approach is twofold: on the one hand, it allows identifying the variables that influence e-commerce sales with an accessible approach. And on the other hand, it can be used to model future sales from the input variables. Results show that e-commerce is linearly dependent on 11 economic and demographic indicators.

***Keywords*—** e-commerce, statistical modeling, regression, empirical research.

I. Introduction

E-commerce embodies a particularly complex system that challenges logistics systems in today's environment. Online shopping is related to a high fragmentation of receivers and deliveries of smaller quantities. The result is that the last mile distribution is the most expensive link of the supply chain [1] [2]. As a result, e-commerce pushes the traditional logistic approaches based on massification [3] [4] [5] in order to remain profitable.

The increasing access to Internet worldwide is allowing consumption through e-commerce in more countries every year. During the past 20 years, although e-commerce in developed countries has been present, it did not grow, as rapidly as expected. It is mostly the past five years, with the adoption of smartphones, tablets and the omnipresence of Internet, that e-commerce has displayed its exponential growth [6].

This paper aims at identifying linear relations between national e-commerce sales and major economic and demographic indicators. This problem has not been explicitly addressed in the literature, but previous works identified important elements on which this paper intends to build.

In 2007, Ho et al. [7] addressed the problem of modeling the growth of e-commerce revenues at a national level, identifying three underlying mechanisms: endogenous, exogenous and a mixed endogenous–exogenous. The study of Ho et al. highlighted the disparity of influences of internal and external drivers on e-commerce growth across countries.

Moreover in 2011, Ho et al. [8] formalized a hybrid growth theory with cross-model inference, which allowed to conclude that "both endogenous variables (e.g. internet user penetration, investment in telecommunication) and exogenous variables (international openness) drive the GDP-normalized level of B2C e-commerce revenues" [8].

Mahmood et al. [9] studied the online shopping behavior of consumers and concluded that e-commerce sales are influenced by two major indicators: namely: (a) economic condition and (b) trust.

The work of Tan et al. [10] indicated that in China, important factors related to B2B e-commerce adoption are: (a) access to computers, (b) lack of internal trust, (c) lack of enterprise-wide information sharing, (d) intolerance towards failure, and (e) incapability of dealing with rapid change.

Meso et al. [11] studied the relationships between information infrastructure and socio-economic development. Their study pointed out that e-commerce adoption is highly related to (a) telephone density, (b) Internet density, (c) GDP, (d) purchasing power parity, among others.

Bejarano [12] proposed forecasting models for e-commerce demand in the U.S.A. and highlighted the difficulty of finding independent variables to predict an unbiased and accurate e-commerce demand.

The following sections of this paper are organized as follows. Section 2 presents the problem statement and the input data. Section 3 presents the proposed approach in which a theoretical background abound linear regression is given and the main steps of the methodology are explained. Section 4 presents the application of the methodology to a case study of 24 countries. And finally Section 5 concludes and opens future research perspectives

II. Problem Statement and Input Data

The problem addressed in this paper consists in identifying the major economic and demographic variables that influence e-commerce sales.

A methodology is proposed to relate the data of e-commerce sales of several countries, to a set of openly available indicators

A. B. Author was with MINES ParisTech at the Center for Robotics (e-mail: apollinairebarme@gmail.com).

S. T. Author is with MINES ParisTech at the Center for Robotics (corresponding author, e-mail: simon.tamayo@mines-paristech.fr, phone: +33140519452)

A. G. Author ia with MINES ParisTech at the Center for Robotics (e-mail: arthur.gaudron@mines-paristech.fr).

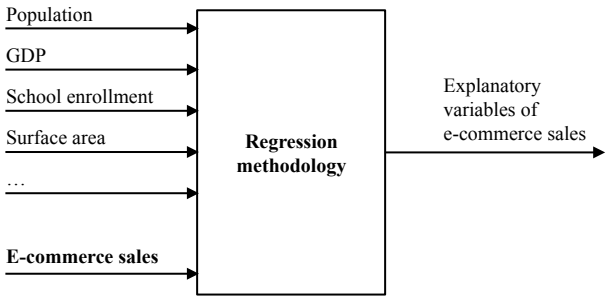

Fig. 1 Overview of the proposed methodology

The presented methodology allows constructing and validating a linear regression model to determine a subset of explanatory variables as shown in Fig 1. As a result, this paper deals with the following research questions:

- Which major variables influence e-commerce sales?
- How to identify these variables?
- Would it be possible to predict the sales of e-commerce in a country from these variables?

The figures of e-commerce sales of 24 countries of were obtained from the Global B2C E-commerce Report 2016 [13].

Economic and demographic data about these 24 countries were obtained in the "Country Profiles Data" in the World Bank Open Data Portal [14]. An initial set of 36 variables was queried. Table I presents a synthesis of these input variables.

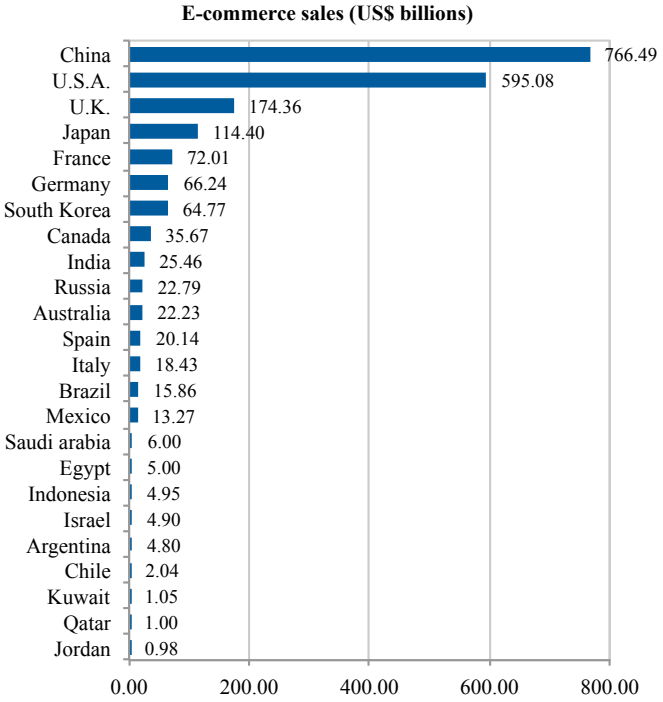

Fig. 2. E-commerce sales (billions of US dollars) in 2015 per country. Source: "Global B2C E-commerce Report 2016" [13]

III. Proposed Approach

The proposed approach is based on a general linear model. This section is presented in three parts. First, the hypotheses of the proposed model are listed. Second, an overview of the theoretical background about linear regression is proposed. And third, the proposed methodology is explained.

TABLE I.
INPUT VARIABLES (SOURCE: WORLD BANK OPEN DATA)

Variable	Unit
1. Population	Millions
2. Population growth	Annual %
3. Surface area	Thousands of sq. km
4. Population density	People per sq. km of land area
5. GNI, Atlas method	Current US$ billions
6. GNI per capita, Atlas method	Current US$
7. GNI, PPP	Current international $ billions
8. GNI per capita, PPP	Current international $
9. Life expectancy at birth	Years
10. Fertility rate	Births per woman
11. Adolescent fertility rate	Births per 1,000 women ages 15-19
12. Mortality rate, under-5	Deaths per 1,000 live births
13. Immunization, measles	% of children ages 12-23 months
14. School enrollment, primary	% gross
15. School enrollment, secondary	% gross
16. Forest area	Thousands of sq. km
17. Terrestrial and marine protected areas	% of total territorial area
18. Improved sanitation facilities	% of population with access
19. Urban population growth	Annual %
20. Energy use	Kg of oil equivalent per capita
21. CO2 emissions	Metric tons per capita
22. Electric power consumption	KWh per capita
23. GDP	Current US$ billions
24. GDP growth	Annual %
25. Inflation, GDP deflator	Annual %
26. Exports of goods and services	% of GDP
27. Imports of goods and services	% of GDP
28. Gross capital formation	% of GDP
29. Time required to start a business	Days
30. Mobile cellular subscriptions	Subscriptions per 100 people
31. Individuals using the Internet	% of population
32. High-technology exports	% of manufactured exports
33. Merchandise trade	% of GDP
34. Net barter terms of trade index	Index 2000 = 100
35. Personal remittances, received	Current US$ millions
36. Foreign direct investment net inflows	Current US$ millions

A. Hypothesis

The proposed methodology is based on a linear regression model. Accordingly, it makes the following hypotheses about the explanatory variables (i.e. indicators) and the dependent variable (i.e. e-commerce sales):

Linearity: the mean of the e-commerce sales result is a linear combination of regression coefficients and the explanatory variables. This assumption is not as restrictive as it seems, because the explanatory variables themselves can be transformed, and in fact multiple copies of the same variable can be added, each one transformed differently [15], leading for example to a polynomial regression function.

Homoscedasticity: the errors in the e-commerce sales results have a constant variance, regardless of the values of the explanatory variables.

Independence of errors: the errors of the e-commerce sales result are uncorrelated with each other.

Lack of multicollinearity: the explanatory variables should not be highly correlated [16]. In order to comply with this

hypothesis a correlation test is performed in the data collection step and the input parameters are narrowed down.

B. Theoretical background

The approach proposed in this paper applies some concepts of statistics that are recalled hereafter.

1) General linear model

A general linear model is used to represent the relationships between e-commerce sales denoted Y (modeled as scalar dependent variable) and a set of explanatory variables denoted X (i.e. the economic and demographic information about each country). Let Rn with its scalar product, its norm and its distance be our reference space. The General Linear Model is represented by the following equation:

$$Y = X\theta + \varepsilon$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1p} \\ 1 & x_{21} & \dots & x_{2p} \\ \dots & \dots & \dots & \dots \\ 1 & x_{n1} & \dots & x_{np} \end{bmatrix} \begin{bmatrix} \theta_0 \\ \theta_1 \\ \dots \\ \theta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_n \end{bmatrix} \quad (1)$$

We assume that: $\boldsymbol{X}$ and $\boldsymbol{Y}$ are both known. Y is a column vector with n entries that contain the dependent variable data. In our case there are 24 countries, therefore n=24. X is matrix containing the explanatory variables data with the first column filled with ones. In order to ensure the existence of a single solution to the linear problem, the number of explanatory variables (p) needs to be less than n-1. The vector θ contains the coefficients of the linear model. The aim of the regression model is to determine θ.

In order to ensure a uniform variance of the expected errors in the model, the Homoscedasticity criterion should be validated (cf. previous section). This criterion is validated if the average expected value of the noises is equal to zero and their variance is constant.

2) Regression

The regression follows the steps shown in (2):

$$\hat{\theta} = \operatorname{arginf} \left\| (Y - X\theta) \right\|^2$$

$$[X] = \left\{ X\theta, \theta \in \Re^{p+1} \right\} \text{ but } [X] \text{ is a finite vector space}$$

$$\text{Hence, } X\hat{\theta} = \operatorname{argmin} d(Y, [X])$$

$$\text{To conclude, due to unicity, } X\hat{\theta} = P_{[X]}(Y)$$

$$\text{Then, the following property is satisfied: } Y\text{-}P_{[X]}(Y) \perp X$$

$$\text{Which leads to } \left\langle Y - P_{[X]}(Y), X \right\rangle = 0$$

$$\text{We deduce that, } X^T(Y - X\hat{\theta}) = 0$$

$$\text{Then, } X^T Y = X^T X\hat{\theta}$$

$$\text{But, } X^T X \in M_{p+1}(\Re) \text{ and } Rank(X^T X) = p+1$$

$$\text{Consequently, } (X^T X)^{-1} \text{ exists}$$

$$\text{To conclude, } \hat{\theta} = (X^T X)^{-1} X^T Y \quad (2)$$

Our aim is to find θ. The linear model hypothesis establishes that $Y = X\theta + \varepsilon$ with ε unknown. To find θ, we look for $X\theta$ the closest to Y. The set X is known, so finding θ or $X\theta$ is equivalent. We then place $[X]$ as the working space (which is a finite vector space), and we minimize the distance between Y and $[X]$, which is obtained for $X\hat{\theta}$, that is, the orthographic projection of Y over $[X]$, as the straight line is that the shortest path between two points.

3) Validation

Many inputs are involved in a regression, some of which are irrelevant. In order to identify a significant subset of explanatory variables, two types of tests are preformed: Student's and Fisher's.

a) Student's Test

The student's test is performed in order to ascertain if a single variable is meaningful to the regression model. This is quantified with the student's *p-value* of each input. For example, in order to verify if θ_1 is a significant parameter, the next test is performed:

$$H_0 : \mathrm{C}^{\mathrm{T}}\theta = 0$$

$$H_1 : \mathrm{C}^{\mathrm{T}}\theta \neq 0$$

$$\mathrm{C}^{\mathrm{T}} = (0,1,...,0) \quad (3)$$

We define T the statistic of the test:

$$T = C^T\theta \left(\hat{\sigma} \left(C^T \left(X^T X \right)^{-1} \right)^{-1/2} \right)^{-1}$$

$$T \text{ follows } St(n-p-1) \quad (4)$$

If $P(|St(n-p-1)|>T) > 0.05$ then we reject H_1 else we reject H_0. In other words, if H_1 is validated, the parameter is significant. $P(|St(n-p-1)|>T)$ is known as the "*p-value*" (for more insight on the theoretical aspects of this test see [16]).

b) Fisher's Test

The fisher's test is performed in order to detect if a set of inputs is significant. It is important to note that the student's test can yield a set of parameters that are significant individually, but that are not relevant as a set. This is why the fisher's test often complements the student's test. For example, in order to verify if the set $\theta_1, \theta_2, \theta_3$ is significant, the next test is performed:

$$H_0 : C^T\theta = 0$$

$$H_1 : H_0 \text{ is false}$$

$$\mathrm{C}^T = \begin{pmatrix} 0 & 1 & -1 & 0 & 0 & \dots & 0 \\ 0 & 0 & 1 & -1 & 0 & \dots & 0 \end{pmatrix} \quad (5)$$

We define T the statistic of the test:

$$T = \theta^T C \left(C^T \left(X^T X \right)^{-1} C \right)^{-1} C^T \theta \hat{\sigma}^{-2}$$

$$T \text{ follows } F(p,\ n-p-1) \quad (6)$$

If $P(|St(n-p-1)|>T) > 0.05$ then we reject H_1 else we reject H_0. As result, the fisher's test returns one single *p-value* result. If this result is less than or equal to *0.05*, the set of parameters is considered to be significant.

C. Proposed methodology

The proposed methodology is divided in three main steps:

(1) Data collection; (2) Regression and selection; and (3) Validation, presented as shown in Fig. 3.

4) Data collection

The presented analysis is based on a sample of 24 countries (*n=24*). In order to be exhaustive, we queried as many parameters as we could. Nonetheless, as indicated previously, the number of explanatory variables needs to be less than *n-1*. As a result only 22 indicators were selected. We measured the correlation between the data and removed the highly correlated indicators. This was done by calculating the correlations matrix and eliminating the correlations higher than 85%. This procedure was repeated iteratively until 22 inputs were obtained.

5) Regression

Once the data is selected, we perform the linear regression following the steps in (2) and we compute the Student's *p-values* for the different parameters, the Fisher's *p-value* for the set and the R^2, this latter, corresponds to the percentage of variances explained, which is a measure of how the model is fitting the observations.

The Fisher's *p-value* is calculated for the entire set of input variables (22). If the *p-value* is above 0.05 the set is globally irrelevant. In this case, we use the Bayesian Information Criterion (BIC), which allows selecting the best-fitted set of variables. The BIC returns a more relevant subset of variables that are fed again to the regression model.

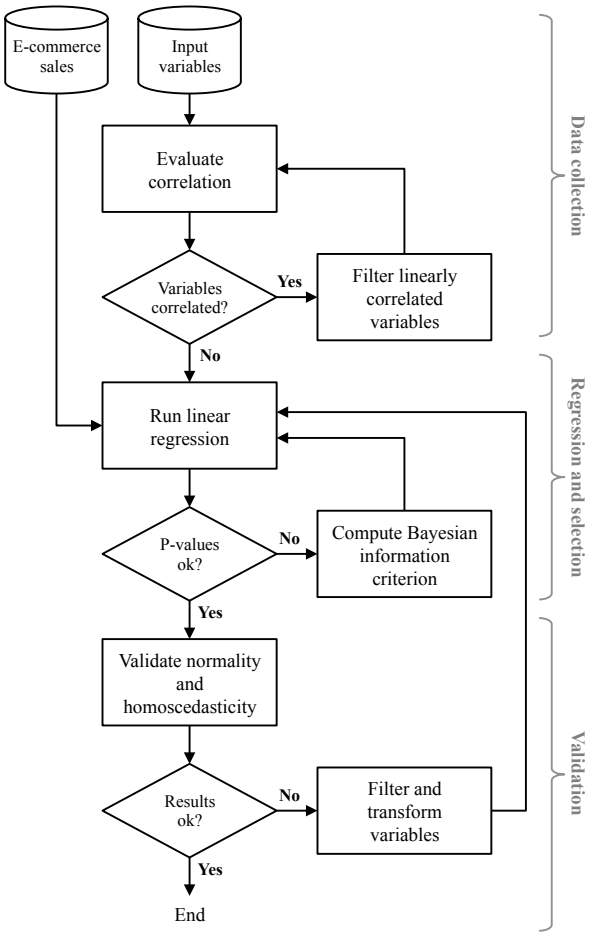

Fig. 3 Detail of the proposed methodology

6) Validation

Once the subset of parameters is validated in terms of *p-values*, three graphs associated with the regression are analyzed:

1. "Residuals versus fitted" shows the positions of the residuals;
2. "Scale location" shows the variances of the residuals;
3. "Residuals vs. leverage" allows detecting aberrant points (if any).

If the variances of the residuals are not uniformly distributed along the axis, the residuals are considered to be heteroscedastic. In this case, the linear regression cannot be applied (cf. hypotheses of the model in section III.A). If such is the case, we perform a Box-Cox transformation on the dependent variable in order to obtain homoscedastic residuals.

The Box-Cox transformation is a power transformation, which determinates the proper power on the variables to ameliorate the regression. After the Box-Cox succeeds, regression is re-applied with the transformed variables.

At the end of the methodology, if all *p-values* are below 0.05 (Fisher's and Student's), the R^2 is close to 1, the epsilons are homoscedastic and normally distributed, the regression is validated and the final subset of explanatory indicators is obtained.

IV. Application to E-Commerce Sales

The application presented in this section was implemented within the "R" software using the libraries "MASS" and "CAR".

A. Data collection

The proposed methodology was applied to a sample of 24 countries. An initial set of 36 explanatory variables was considered (cf. section II). After filtering the input data, significant correlations were identified, and the initial set was reduced to 22 variables on the grounds of the correlation coefficients.

B. Regression

A linear regression was performed following the steps shown in (2) and we obtained an adjusted R2 equal to 0.4654 (which is not yet satisfactory) and a Fisher's p-value of 0.5231. Table II presents a summary of the regression results for the 22 variables.

The Fisher's *p-value* result (higher than 0.05) indicates that this set of variables is not significant (cf. section III.B.2). As result, we use the BIC algorithm in order to filter the input variables. After running the BIC procedure, we obtained a subset of 14 variables, presented in table III.

The Bayesian Criterion indicates that these variables should be explanatory inputs for the sales observation. The BIC column in Table III indicates that the 14 variables are relevant (i.e. the BIC result of each variable is greater than that of "none").

TABLE II.
RESULTS OF THE FIRST REGRESSION

	Estimate	Std. Error	Pr(>|t|)
(Intercept)	6.172e+03	2.159e+06	0.998
Population	1.672e+02	3.637e+02	0.726
Population growth	1.683e+05	2.131e+05	0.574
Surface area	-1.758e+01	1.827e+01	0.512
Population density	-1.205e+03	1.178e+03	0.493
GNI per capita	7.935e+00	8.269e+03	0.999
Fertility rate	-2.645e+05	3.002e+05	0.540
Adolescent fertility	-4.510e+03	7.103e+03	0.640
Immunization measles	7.421e+03	1.074e+04	0.615
School enrollment prim.	7.691e+03	2.295e+04	0.794
School enrollment sec.	-3.092e+03	1.011e+04	0.811
Terr. and marine pro.	-6.950e+03	7.973e+03	0.544
GDP	1.565e-02	3.430e-02	0.727
GDP growth	2.524e+04	8.458e+04	0.815
Inflation	8.252e+02	8.159e+03	0.936
Imports of goods	-9.231e+03	2.014e+04	0.726
Gross capital formation	-9.278e+03	1.437e+04	0.635
Time to start a business	-2.460e+03	6.188e+03	0.759
Mobile cellphone subscriptions	2.637e+03	4.145e+03	0.639
Individuals using the internet	-4.983e+03	1.159e+04	0.742
High technology exports	2.006e+04	1.610e+04	0.431
Merchandise trade	-5.175e+02	9.343e+03	0.965
Net barter term	-1.376e+03	3.787e+03	0.778

TABLE III.
RESULTS OF THE BIC SELECTION

	Sum of Sq.	RSS	BIC
None		2.7841e+10	547.29
Gross capital formation	5.1831e+09	3.3024e+10	548.29
Population	7.6372e+09	3.5478e+10	550.01
Mobile cellphone subscriptions	1.2806e+10	4.0647e+10	553.28
Fertility rate	1.4667e+10	4.2508e+10	554.35
Population growth	1.6154e+10	4.3995e+10	555.18
Import of goods	1.6181e+10	4.4022e+10	555.19
Individuals using the internet	1.7300e+10	4.5141e+10	555.79
Immunization measles	1.9147e+10	4.6988e+10	556.76
Adolescent fertility	2.3058e+10	5.0899e+10	558.68
Terrestrial and marine protected	2.4698e+10	5.2539e+10	559.44
Surface area	3.3475e+10	6.1316e+10	563.14
High technology exports	4.7518e+10	7.5359e+10	568.09
Population density	4.7832e+10	7.5673e+10	568.19
GDP	8.1840e+10	1.0968e+11	577.10

Once this new set is established we run once more the linear regression and we obtained an adjusted R^2 equal to 0.9139 and a Fisher's *p-value* of 6.306e-05. The results of this second regression are shown in Table IV.

The Fisher's *p-value* is below 0.05, as a result, the sample is significant. The R^2 is close to 1, nonetheless some variables have *p-values* bigger than 0.05 ("Population", "Gross capital formation" and "mobile cellular subscriptions"). We removed those variables one by one verifying that the resulting set had a valid fisher's *p-value*, and we obtained a new subset of 11 indicators.

TABLE IV.
RESULTS OF THE SECOND REGRESSION

	Estimate	Std. Error	Pr(>|t|)
(Intercept)	2.748e+05	3.516e+05	0.454586
Population	1.238e+02	7.877e+01	0.150571
Population growth	1.206e+05	5.276e+04	0.048152
Surface area	-1.743e+01	5.298e+00	0.009384
Population density	-8.596e+02	2.186e+02	0.003446
Fertility rate	-1.518e+05	6.972e+04	0.057413
Adolescent fertility	-4.945e+03	1.811e+03	0.023218
Immunization measles	6.920e+03	2.782e+03	0.034541
Terrestrial and marine protected	-5.989e+03	2.119e+03	0.019865
GDP	2.435e-02	4.734e-03	0.000608
Imports of goods	-6.371e+03	2.786e+03	0.048002
Gross capital formation	-4.446e+03	3.435e+03	0.227746
Mobile cellphone subscriptions	9.998e+02	4.914e+02	0.072391
Individuals using the internet	-6.251e+03	2.643e+03	0.042264
High technology exports	1.716e+04	4.378e+03	0.003515

A final regression was run with this final subset and we obtained an adjusted R^2 equal to 0.8915 and a Fisher's *p-value* of 8.296e-06, which are both satisfactory. The results of this third regression are shown in Table V.

TABLE V.
RESULTS OF THE FINAL REGRESSION

	Estimate	Std. Error	Pr(>|t|)
(Intercept)	1.138e+05	2.329e+05	0.63382
Population growth	8.968e+04	2.904e+04	0.00940
Surface area	-1.229e+01	4.957e+00	0.02897
Population density	-6.228e+02	1.825e+02	0.00516
Fertility rate	-1.087e+05	4.149e+04	0.02234
Adolescent fertility rate	-4.062e+03	1.158e+03	0.00432
Immunization measles	7.285e+03	2.577e+03	0.01524
Terrestrial and marine protected	-4.148e+03	1.744e+03	0.03489
GDP	2.889e-02	4.506e-03	3.35e-05
Imports of goods	-5.082e+03	2.193e+03	0.03895
Individuals using the internet	-5.770e+03	1.308e+03	0.00085
High technology exports	1.266e+04	3.060e+03	0.00138

All the variables have their *p-values* below 0.05, the Fisher's *p-value* is adequate and the R^2 is more than sufficient. The summary suggests that the regression is meaningful. We now analyze the graphs to validate these results.

C. Validation

The scale location graph (Fig. 4) shows the homogeneity between the variances of the residuals. That is, the epsilons are uniformly spread along the axis, which indicates that they are homoscedastic. Moreover, the Breush-Pagan test [17] reinforces the hypothesis of homoscedasticity of the residuals.

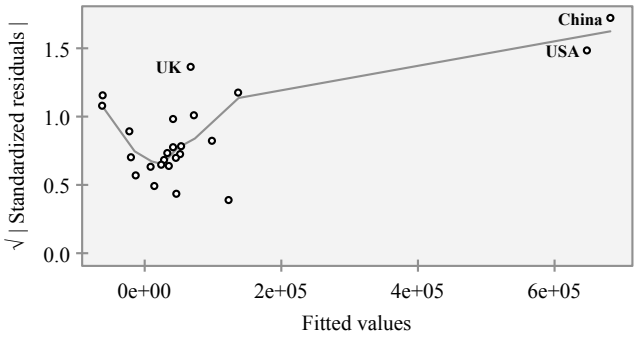

Fig. 4 Scale location

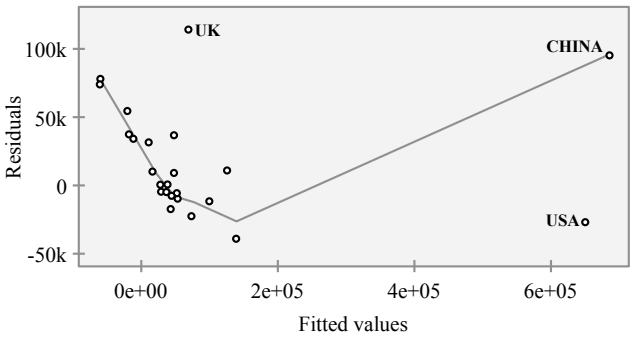

Fig. 5 Residuals versus fitted values

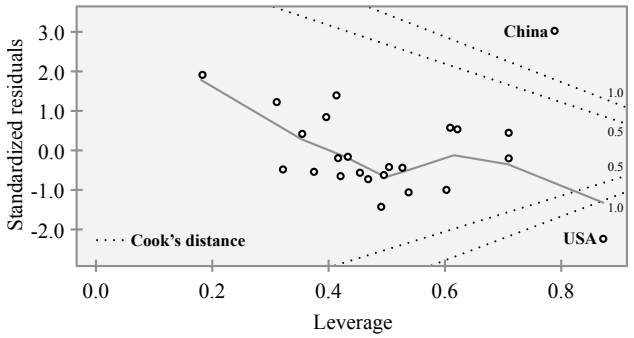

Fig. 6 Residuals versus leverage

Fig.5 presents the location of the residuals. This graph validates the linearity of the model, with the exception of the two points in the right. That is, the observations of the sales in China and U.S.A.. These points degrade the regression results. From a practical point of view, this means that in these two countries e-commerce sales follow a different logic than that of the rest of the sample.

As mentioned in the theoretical background, the "residuals versus leverage" (Fig. 6) graph indicates atypical points (those that are behind the Cook's distance boundary [15] [16]). Fig. 6 points out that the regression would be improved if the first and the fourth observations (i.e. China and U.S.A.) were removed. Nonetheless, because of the size of the sample (only 24 observations) it is not possible to remove these points. It is important to note that the R^2 and the *p-values* indicate that the regression is relevant event with these two abnormal observations.

Moreover, in Figures 4 and 5, U.K. appears to be a remote point. Yet, in Fig. 6 the cook's distance shows that the U.K. is within the regular boundaries. As result, U.K. is considered to be an extreme point that remains relevant to the regression.

D.Discussion

36 variables were included in our initial data set. Due to this large number of parameters, we use the correlation between them to erase the correlated entries. Once the set was reduced to 22 variables, we proceeded to the regression, the result was unequivocal: the set was still not reliable. The BIC algorithm allowed to the narrow down the set to 14 variables. The regression was run again and three variables with high *p-values*, were also removed. Finally, we obtained a final set of 11. A last regression indicated that this was a valid set (adjusted R^2 = 0.8915 and Fisher's *p-value* = 8.296e-06).

The regression indicates that the indicators with the higher influence on e-commerce sales are: (1) GDP, (2) Individuals using Internet, (3) High technology exports and (4) Population density. This is observed from the student's *p-values* results in Table V. It is important to note, that these indicators were also identified as influential parameters when assessing B2B e-commerce adoption in the works of Ho et al. [8] and Meso et al. [11]. GDP is the most significant indicator the lowest *p-value* of 3.35e-05. Fig. 7 overlaps the graphs of E-commerce sales and GDP for the studied countries, in which the linear correspondence is remarkable.

Other interesting variables, such as "Adolescent fertility", "Immunization measles" and "Terrestrial and marine protected" are present in the regression, these are surprising results, but that may quantify in a certain way the development of the countries.

It is to note that regression identifies China and U.S.A. as aberrant points (cf. Fig. 4 and Fig. 5). However it is not surprising that the homelands of Amazon and Alibaba have a behavior different than the average.

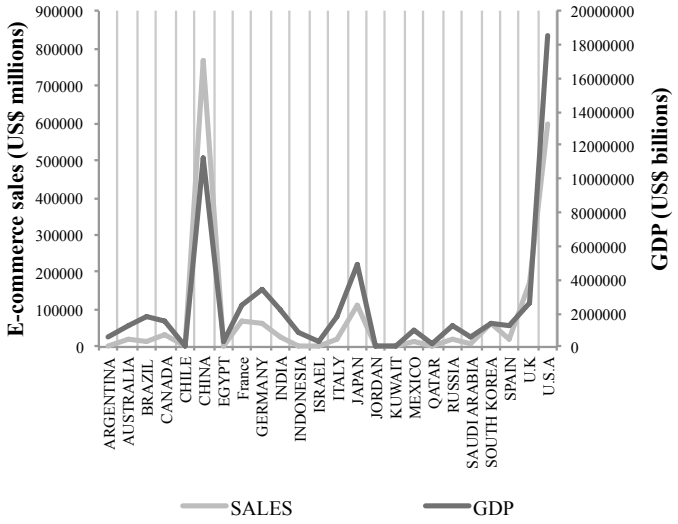

Fig.7 Overlap of the e-commerce sales and the GDP

V. CONCLUSION

Regression is a statistical tool that estimates the relationship between a dependent observation and a set of explanatory variables. This paper proposes a methodology to relate e-

commerce sales to economic and demographic indicators. This paper addressed three research questions that are reviewed hereafter.

Which major variables influence e-commerce sales? A set of 11 explanatory variables has been established. It includes the following indicators in order of significance: (1) GDP, (2) Individuals using Internet, (3) High technology exports, (4) Adolescent fertility rate, (5) Population density, (6) Population growth, (7) Immunization, measles, (8) Fertility rate, (9) Surface area, (10) Terrestrial and marine protected areas and (11) Imports of goods and services.

How to identify these variables? A methodology was proposed in order to validate the significance of a set of input indicators. First the indicators are filtered on the grounds of linear correlation between them. Then a subset of independent indicators is fed into the linear regression model, which in turn, iteratively narrows these indicators in order to validate the hypotheses of linearity, independence and homoscedasticity. The methodology is presented in thee main steps: Data collection, Regression and selection and Validation.

Would it be possible to predict the aggregate sales of e-commerce in a certain country from these variables? A low *p-value* and high R^2 results indicated that the resulting model is adequate. However, the input set of observations is still very limited: only 24 countries form the set of dependent variables. A sample size this small cannot yield relevant prediction results. In order to assess the predictive capabilities of the methodology, the model was build with the data of 23 countries, and the sales the remaining one were predicted. This was done with different combinations of countries, and the average error was about 40%.

A future version of this approach should take into account more sales observations and more input indicators. Also it would be interesting to perform similar regressions data of consecutive years in order to deduce parameters linked to e-commerce growth, such as the arrival of a big player, a significant improvement in the access to the Internet, etc.

The proposed approach has two important limitations. First, the linear model hypothesis supposes that the influence of the explanatory variables is uniform regardless of the country. For example, it admits that the GDP has as much influence on e-commerce sales in China as in Jordan, yet this assumption can be challenged. Second, the input data set is far from ideal: on the one hand, the sample size is small, which limits the number of indicators that can be considered (cf. section III) and on the other hand, there is no measure of the quality of the data.

ACKNOWLEDGMENT

This work is supported by the Urban Logistics Chair at MINES ParisTech, sponsored by ADEME (Agency for the Environment and Energy Management), La Poste Group, Mairie de Paris (Paris City Hall), Pomona Group and RENAULT.

REFERENCES

[1] J. Allen, M. Piecyk, M. Piotrowska, F. McLeod, T. Cherrett, K. Ghali, T. Nguyen, T. Bektas, O. Bates, A. Friday, S. Wise, and M. Austwick, "Understanding the impact of e-commerce on last-mile light goods vehicle activity in urban areas: The case of London," Transp. Res. Part D Transp. Environ., 2017.

[2] R. Gevaers, E. Van de Voorde, and T. Vanelslander, "Cost Modelling and Simulation of Last-mile Characteristics in an Innovative B2C Supply Chain Environment with Implications on Urban Areas and Cities," Procedia - Soc. Behav. Sci., vol. 125, pp. 398–411, 2014.

[3] Y. Yu, X. Wang, R. Y. Zhong, and G. Q. Huang, "E-commerce Logistics in Supply Chain Management: Practice Perspective," Procedia CIRP, vol. 52, pp. 179–185, 2016.

[4] S. Čavoški and A. Marković, "Agent-based modelling and simulation in the analysis of customer behaviour on B2C e-commerce sites," J. Simul., Oct. 2016.

[5] V. Jain, S. Wadhwa, and S. G. Deshmukh, "e-Commerce and supply chains: Modelling of dynamics through fuzzy enhanced high level petri net," Sadhana - Acad. Proc. Eng. Sci., vol. 30, no. 2–3, pp. 403–429, 2005.

[6] S. J. Winter, "The rise of cyberinfrastructure and grand challenges for eCommerce," Inf. Syst. E-bus. Manag., vol. 10, no. 3, pp. 279–293, 2012.

[7] S. C. Ho, R. J. Kauffman, and T. P. Liang, "A growth theory perspective on B2C e-commerce growth in Europe: An exploratory study," Electron. Commer. Res. Appl., vol. 6, no. 3, pp. 237–259, 2007.

[8] S. C. Ho, R. J. Kauffman, and T. P. Liang, "Internet-based selling technology and e-commerce growth: A hybrid growth theory approach with cross-model inference," Inf. Technol. Manag., vol. 12, no. 4, pp. 409–429, 2011.

[9] M. A. Mahmood, K. Bagchi, and T. C. Ford, "On-line shopping behavior: Cross-country empirical research," Int. J. Electron. Commer., vol. 9, no. 1, pp. 9–30, 2004.

[10] J. Tan, K. Tyler, and A. Manica, "Business-to-business adoption of eCommerce in China," Inf. Manag., vol. 44, no. 3, pp. 332–351, Apr. 2007.

[11] P. Meso, P. Musa, D. Straub, and V. Mbarika, "Information infrastructure, governance, and socio-economic development in developing countries," Eur. J. Inf. Syst., vol. 18, no. 1, pp. 52–65, Feb. 2009.

[12] A. Bejerano, "Challenges of forecasting demand for e_commerce," Arkansas, U, 2016.

[13] Ecommerce Foundation, "Global B2C E-commerce Report 2016," Amsterdam - the Netherlands, 2016.

[14] World Bank Group, "Country Profiles Data," Washington, USA, 2016.

[15] G. A. F. Seber and A. J. Lee, Linear Regression Analysis, 2nd ed. Wiley, 2003.

[16] J.-M. Azais and J.-M. Bardet, Le modèle linéaire par l'exemple. Dunod, 2006.

[17] T. E. Society, "A Simple Test for Heteroscedasticity and Random Coefficient Variation Author (s): T . S . Breusch and A . R . Pagan Published by : The Econometric Society Stable URL : http://www.jstor.org/stable/1911963," vol. 47, no. 5, pp. 1287–1294, 2010.