Going Beyond GDP to Nowcast Well-Being Using Retail Market Data

Riccardo Guidotti[1,2], Michele Coscia[3], Dino Pedreschi[2]
and Diego Pennacchioli[1]

[1] KDDLab ISTI CNR, Via G. Moruzzi, 1, Pisa, IT {name.surname}@isti.cnr.it
[2] KDDLab CS Dept. Univ. of Pisa, L. B. Pontecorvo, 3, Pisa, IT
{name.surname}@di.unipi.it
[3] CID - HKS, 79 JFK St. Cambridge MA, US michele_coscia@hks.harvard.edu

Abstract. One of the most used measures of the economic health of a nation is the Gross Domestic Product (GDP): the market value of all officially recognized final goods and services produced within a country in a given period of time. GDP, prosperity and well-being of the citizens of a country have been shown to be highly correlated. However, GDP is an imperfect measure in many respects. GDP usually takes a lot of time to be estimated and arguably the well-being of the people is not quantifiable simply by the market value of the products available to them. In this paper we use a quantification of the average sophistication of satisfied needs of a population as an alternative to GDP. We show that this quantification can be calculated more easily than GDP and it is a very promising predictor of the GDP value, anticipating its estimation by six months. The measure is arguably a more multifaceted evaluation of the well-being of the population, as it tells us more about how people are satisfying their needs. Our study is based on a large dataset of retail micro transactions happening across the Italian territory.

1 Introduction

Objectively estimating a country's prosperity is a fundamental task for modern society. We need to have a test to understand which socio-economic and political solutions are working well for society and which ones are not. One such test is the estimation of the Gross Domestic Product, or GDP. GDP is defined as the market value of all officially recognized final goods and services produced within a country in a given period of time. The idea of GDP is to capture the average prosperity that is accessible to people living in a specific region.

No prosperity test is perfect, so it comes as no surprise to reveal that GDP is not perfect either. GDP has been harshly criticised for several reasons [5]. We focus on two of these reasons. First: GDP is not an easy measure to estimate. It takes time to evaluate the values of produced goods and services, as to evaluate them they first have to be produced and consumed. Second: GDP does not accurately capture the well-being of the people. For instance income inequality skews the richness distribution, making the per capita GDP uninteresting, because it does not describe the majority of the population any more. Moreover,

arguably it is not possible to quantify well-being just with the number of dollars in someone's pocket: she might have dreams, aspirations and sophisticated needs that bear little to no correlation with the status of her wallet.

In this paper we propose a solution to both shortcomings of GDP. We introduce a new measure to test the well-being of a country. The proposed measure is the average sophistication of the satisfiable needs of a population. We are able to estimate such measure by connecting products sold in the country to the customers buying them in significant quantities, generating a customer-product bipartite network. The sophistication measure is created by recursively correcting the degree of each customer in the network. Customers are sophisticated if they purchase sophisticated products, and products are sophisticated if they are bought by sophisticated customers. Once this recursive correction converges, the aggregated sophistication level of the network is our well-being estimation.

The average sophistication of the satisfiable needs of a population is a good test of a country's prosperity as it addresses the two issues of GDP we discussed. First, it shows a high correlation with the GDP of the country, when shifting the GDP by two quarters. The average sophistication of the bipartite network is an effective nowcasting of the GDP, making it a promising predictor of the GDP value the statistical office will release after six months. Second, our measure is by design an estimation of the sophistication of the needs satisfied by the population. It is more in line with a real well-being measure, because it detaches itself from the mere quantity of money circulating in the country and focuses closely on the real dynamics of the population's everyday life.

The analysis we present is based on a dataset coming from a large retail company in Italy. The company operates ~ 120 shops in the West Coast in Italy. It serves millions of customers every year, of which a large majority is identifiable through fidelity cards. We analyze all items sold from January 2007 to June 2014. We connect each customer to all items she purchased during the observation period, reconstructing 30 quarterly bipartite customer-product networks. For each network, we quantify the average sophistication of the customers and we test its correlation with GDP, for different temporal shift values.

2 Related Work

Nowcasting is a promising field of research to resolve the delay issues of GDP. Nowcasting has been successfully combined with the analysis of large datasets of human activities. Two famous examples are Google Flu trends [26] and the prediction of automobile sales [4]. Social media data has been used to nowcast employment status and shocks [25] [20]. Such studies are not exempt from criticisms: [18] proved that nowcasting with Google queries alone is not enough and the data must be integrated with other models. Nowcasting has been already applied to GDP too [10], however the developed model uses a statistical approach that is intractable for a high number of variables, thus affecting the quality of results. Other examples can be found focusing on the Eurozone [7], or on different targets such as poverty risk [22] and income distribution [19].

Our proposal of doing GDP nowcasting using retail data is based on the recent branch of research that considers markets as self-organizing complex systems. In [13], authors model the global export market as a bipartite network, connecting the countries with the products they export. Such structure is able to predict long-term GDP growth of a country. This usage of complex networks has been replicated both at the macro economy level [2] and at the micro level of retail [3]. At this level, in previous work we showed that the complex system perspective still yields an interesting description of the retail dynamics [23]. We defined a measure of product and customer sophistication and we showed its power to explain the distance travelled by customers to buy the products they need [24], and even their profitability for the shop [12]. In this work, we borrow these indicators and we use them to tackle the problem of nowcasting GDP. An alternative methodology uses electronic payment data [8]. However in this case the only issue addressed is the timing issue, but no attempt is made into making the measure more representative of the satisfaction of people's needs.

The critiques to GDP we mentioned have resulted in the proliferation of alternative well-being indicators. We mention the Index of Sustainable Economic Welfare (ISEW), the Genuine Progress Indicator (GPI) [16] and the Human Development Index (HDI)[1]. A more in depth review about well-being alternatives is provided in [14]. These indicators are designed to correct some shortcomings of GDP, namely incorporating sustainability and social cost. However, they are still affected by long delays between measurements and evaluation. They are also affected by other criticisms: for instance, GPI includes a list of adjustment items that is considered inconsistent and somewhat arbitrary. Corrections have been developed [17], but so far there is no final reason to prefer them to GDP and thus we decide to adhere to the standard and we consider only the GDP measure, and we remark that no alternative has addressed the two mentioned issues of GDP in a universally recognized satisfactory way.

3 Data

Our analysis is based on real world data about customer behaviour. The dataset we used is the retail market data of one of the largest Italian retail distribution companies. The dataset has been already presented in previous works ([12] [24]) and we refer to those publications for an in-depth description of our cleaning strategy. We report here when we perform different operations.

The dataset contains retail market data in a time window spanning from Jan 1st, 2007 to June, 30th 2014. The active and recognizable customers are $\sim 1M$. The stores of the company cover the West Coast of Italy. We aggregated the items sold using the Segment classification in the supermarket's marketing hierarchy. We end up with $\sim 4,500$ segments, to which we refer as products.

At this point we need to define the time granularity of our observation period. We choose to use a quarterly aggregation mainly because we want to compare

[1] `http://hdr.undp.org/en/statistics/hdi`

our results with GDP, and GDP assumes a better relevance in a quarterly aggregation. For each quarter, we have $\sim 500k$ active customers.

Since our objective is to establish a correlation between the supermarket data and the Gross Domestic Product of Italy, we need a reliable data source for GDP. We rely on the Italian National Bureau of Statistic ISTAT. ISTAT publishes quarterly reports about the status of the Italian country under several aspects, including the official GDP estimation. ISTAT is a public organization and its estimates are the official data used by the Italian central government. We downloaded the GDP data from the ISTAT website[2].

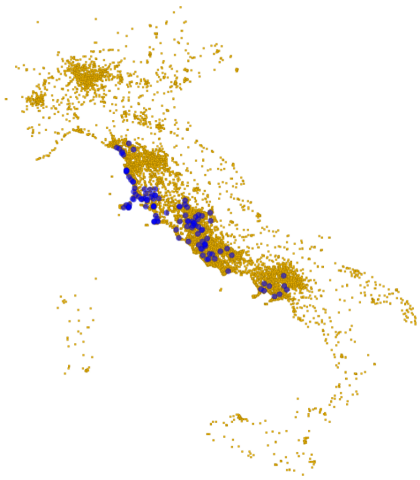

Fig. 1: The geographical distribution of observed customers (yellow dots) and shops (blue dots) in the territory of Italy.

Figure 1 shows that the observed customers cover the entire territory of Italy. However, the shop distribution is not homogeneous. Shops are located in a few Italian regions. Therefore, the coverage of these regions is much more significant, while customers from other regions usually shop only during vacation periods in these regions. Our analysis is performed on national GDP data, because regional GDP data is disclosed only with a yearly aggregation. However, the correlation between national GDP and the aggregated GDP of our observed regions (Tuscany, Lazio and Campania) during our observation period is 0.95 ($p < 0.001$). This is because Italy has a high variation on the North-South axis, which we cover, while the West-East variation, which we cannot cover, is very low.

[2] `http://dati.istat.it/Index.aspx?lang=en&themetreeid=91`, date of last access: September 23rd, 2015

4 Methodology

In this section we present the methodology implemented for the paper. First, we present the algorithm we use to estimate the measure of sophistication (Section 4.1). Second, we discuss the seasonality issues (Section 4.2).

4.1 Sophistication

The sophistication index is used to objectively quantify the sophistication level of the needs of the customers buying products. We introduced the sophistication index in [24], which is an adaptation from [13], necessary to scale up to large datasets. We briefly report here how to compute the customer sophistication index, and we refer to the cited papers for a more in-depth explanation.

The starting point is a matrix with customers on rows and products on the columns. This matrix is generated for each quarter of each year of observation. Each cell contains the number of items purchased by the customer of the product in a given quarter (e.g. Q1 of 2007, Q2 of 2007 and so on). We then have 30 of such matrices. The matrices are already very sparse, with an average fill of 1.4% (ranging from 33 to 37 million non zero values). Our aim is to increase the robustness of these structures, by constructing a bipartite network connecting customers exclusively to the subset of products they purchase in significant quantities. Figure 2 provides a simple depiction of the output bipartite network.

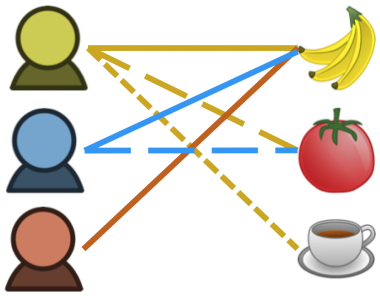

Fig. 2: The resulting bipartite network connecting customers to the products they buy in significant quantities.

To filter the edges, we calculate the Revealed Comparative Advantage (RCA, known as Lift in data mining [9]) of each product-customer cell [1], following [13]. Given a product p_i and a customer c_j, the RCA of the couple is defined as follows:

$$RCA(p_i, c_j) = \frac{X(p_i, c_j)}{X(p_*, c_j)} \left(\frac{X(p_i, c_*)}{X(p_*, c_*)} \right)^{-1},$$

where $X(p_i, c_j)$ is the number of p_i bought by c_j, $X(p_*, c_j)$ is the number of products bought by c_j, $X(p_i, c_*)$ is the total number of times p_i has been sold and $X(p_*, c_*)$ is the total number of products sold. RCA takes values from 0 (when $X(p_i, c_j) = 0$, i.e. customer c_j never bought a single instance of product p_i) to $+\infty$. When $RCA(p_i, c_j) = 1$, it means that $X(p_i, c_j)$ is exactly the expected value under the assumption of statistical independence, i.e. the connection between customer c_j and product p_i has the expected weight. If $RCA(p_i, c_j) < 1$ it means that the customer c_j purchased the product p_i less than expected, and vice-versa. Therefore, we keep an edge in the bipartite network iff its corresponding RCA is larger than 1. Note that most edges were already robust. When filtering out the edges, we keep 93% of the original connections.

The customer sophistication is directly proportional to the customer's degree in the bipartite network, i.e. with the number of different products she buys. Differently from previous works [24] that used the traditional economic complexity algoritm [13], in this work we use the Cristelli formulation of economic complexity [6]. Note that the two measures are highly correlated. Therefore, in the context of this paper, there is no reason to prefer one measure over the other, and we make the choice of using only one for clarity and readability.

Consider our bipartite network $G = (C, P, E)$ described by the adjacency matrix $M^{|C|\times|P|}$, where C are customers and P are products. Let c and p be two ranking vectors to indicate how much a C-node is linked to the most linked P-nodes and, similarly, P-nodes to C-nodes. It is expected that the most linked C-nodes connected to nodes with high p_j score have an high value of c_i, while the most linked P-nodes connected to nodes with high c_i score have an high value of p_j. This corresponds to a flow among nodes of the bipartite graph where the rank of a C-node enhances the rank of the P-node to which is connected and vice-versa. Starting from $i \in C$, the unbiased probability of transition from i to any of its linked P-nodes is the inverse of its degree $c_i^{(0)} = \frac{1}{k_i}$, where k_i is the degree of node i. P-nodes have a corresponding probability of $p_j^{(0)} = \frac{1}{k_j}$. Let n be the iteration index. The sophistication is defined as:

$$c_i^{(n)} = \sum_{j=1}^{|P|} \frac{1}{k_j} M_{ij} p_j^{(n-1)} \forall i \quad p_j^{(n)} = \sum_{i=1}^{|C|} \frac{1}{k_i} M_{ij} c_i^{(n-1)} \forall j$$

These rules can be rewritten as a matrix-vector multiplication

$$c = \bar{M} p \quad p = \bar{M}^T c \tag{1}$$

where $\bar{M}$ is the weighted adjacency matrix. So, like previously we have

$$c^{(n)} = \bar{M}\bar{M}^T c^{(n-1)} \quad p^{(n)} = \bar{M}^T \bar{M} p^{(n-1)}$$

$$c^{(n)} = \mathcal{C} c^{(n-1)} \quad p^{(n)} = \mathcal{P} p^{(n-1)}$$

where $\mathcal{C}^{(|C|\times|C|)} = \bar{M}\bar{M}^T$ and $\mathcal{P}^{(|P|\times|P|)} = \bar{M}^T\bar{M}$ are related to $x^{(n)} = Ax^{(n-1)}$. This makes sophistication solvable using the power iteration method

(and it is proof of convergence). Note that this procedure is equivalent to the HITS ranking algorithm, as proved in [11].

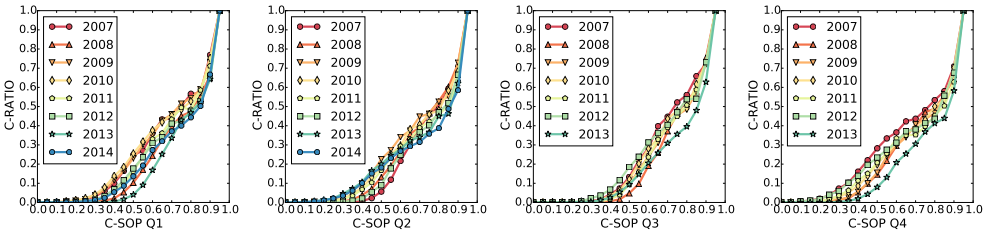

Fig. 3: The customer sophistication distributions per quarter and per year. Each plot reports the probability (y axis) of a customer to have a given sophistication value (x axis), from quarter 1 to quarter 4 (left to right) for each year.

At the end of our procedure, we have a value of customer and product sophistication for each customer for each quarter. For the rest of the section we focus on customer sophistication for space reasons. Each customer is associated with a timeline of 30 different sophistications. The overall sophistication is normalized to take values between 0 and 1. Figure 3 shows the distribution of the customer sophistication per quarter and per year. We chose to aggregate the visualization by quarter because the same quarters are similar across years but different within years, due to seasonal effects.

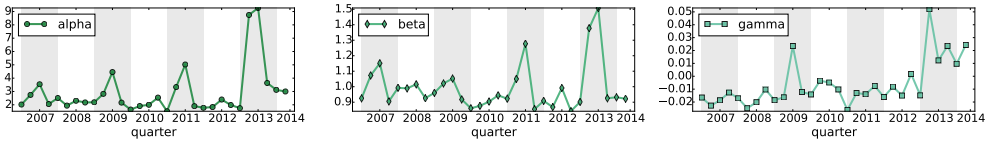

Fig. 4: The different values taken by the fit parameters across the observation period for the sophistication distribution.

Figure 3 shows that the sophistication distribution is highly skewed. We expect it to be an exponential function: by definition the vast majority of the population is unsophisticated and highly sophisticated individuals are an elite. The fit function cannot be a power-law because the different levels of sophistication for least to most sophisticated do not span a sufficiently high number of orders of magnitude. We fitted a function of the form $f(x) = \gamma + \beta \times \alpha^x$ for each quarterly snapshot of our bipartite networks. Figure 4 reports the evolution of the fit parameters α, β and γ. The figure shows that the fit function is mostly stable over time. The fits have been performed using ordinary least squares regression.

SOP Rank	Product
1	Cosmetics
2	Underwear for men
3	Furniture
4	Multimedia services
5	Toys
...	...
-5	Fresh Cheese
-4	Red Meat
-3	Spaghetti
-2	Bananas
-1	Short Pasta

Table 1: The most and least sophisticated products in our dataset.

To prove the quality of our sophistication measure in capturing need sophistication, we report in Table 1 a list of the top and bottom sophisticated products, calculated aggregating data from all customers. Top sophisticated products are non daily needed products and are usually non-food. The least complex products are food items. Being Italian data, pasta is the most basic product.

4.2 Seasonality

Both GDP and the behavior of customers in the retail market are affected by seasonality. Different periods of the year are associated with different economic activities. This is particularly true for Italy in some instances: during the month of August, Italian productive activities come to an almost complete halt, and the country hosts its peak tourist population. The number and variety of products available in the supermarket fluctuates too, with more fruit and vegetables available in different months, or with Christmas season and subsequent sale shocks.

A number of techniques have been developed to deal with seasonal changes in GDP. One of the most popular seasonal adjustments is done through the X-13-Arima method, developed by the U.S. Census Bureau [21]. However, we are unable to use this methodology for two reasons. First, it requires an observation period longer than the one we are able to provide in this paper. Second, the methodologies present in literature are all fine-tuned to specific phenomena that are not comparable to the shopping patterns we are observing. Thus we cannot apply them to our sophistication timelines. Given that we are not able to make a seasonal adjustment for the sophistication, we chose to not seasonally adjust GDP too. We acknowledge this as a limitation of our study and we leave the development of a seasonal adjustment for sophistication as a future work.

5 Experiments

In this section we test the relation between the statistical properties of the bipartite networks generated with our methodology and the GDP values of the

country. We first show the evolution of aggregated measures of expenditure, number of items, degree and sophistication along our observation period. We then test the correlation with GDP, with various temporal shifts to highlight the potential predictive power of some of these measures.

Before showing the timelines, we describe our approach for the aggregation of the properties of customers. The behavior of customers is highly differentiated. We already shown that the sophistication distribution is highly skewed and best represented as an exponential function. The expenditure and the number of items purchased present a skewed distribution among customers: few customers spend high quantities of money and buy many items, many customers spend little quantities of money and buy few items. For this reason, we cannot aggregate these measures using the average over the entire distribution, as it is not well-behaved for skewed values. To select the data we use the inter-quantile range, the measure of spread from the first to the third quantile. In practice, we trim the outliers out of the aggregation and then we compute the average, the Inter-Quartile Mean, or "IQM". The IQM is calculated as follows:

$$x_{\mathrm{IQM}} = \frac{2}{n} \sum_{i=\frac{n}{4}+1}^{\frac{3n}{4}} x_i$$

assuming n sorted values.

Also note that all the timelines we present have been normalized. All variables take values between zero and one, where zero represents the minimum value observed and one the maximum. As for the notation used, in the text and in the captions of the figures we use the abbreviations reported in Table 2.

Abbreviation	Description
IQM	Inter-Quartile Mean.
GDP	Gross Domestic Product.
EXP	IQM of the total expenditure per customer.
PUR	IQM of the total number of items purchased per customer.
C-DEG	IQM of the number of products purchased in significant quantities (i.e. the bipartite network degree) per customer.
P-DEG	IQM of the number of customers purchasing the product in significant quantities (i.e. the bipartite network degree).
C-SOP	IQM of the sophistication per customer.
P-SOP	IQM of the sophistication per product.

Table 2: The abbreviations for the measures used in the experiment section.

The first relation we discuss is between GDP and the most basic customer variables. Figure 5 depicts the relation between GDP and the IQM expenditure (left), and GDP and IQM of number of items purchased (right). Besides the obvious seasonal fluctuation, we can see that the two measures are failing to

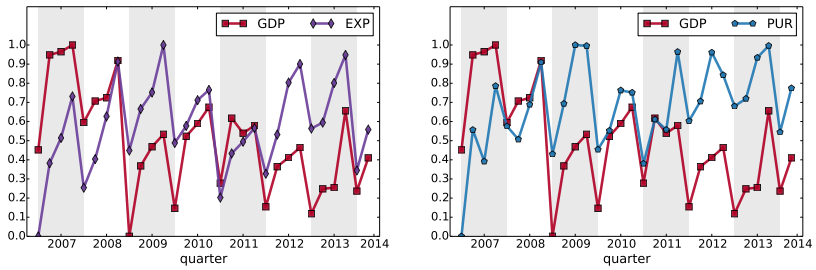

Fig. 5: The relation between GDP and IQM customer expenditure (left) and IQM number of items purchased (right).

capture the overall GDP dynamics. GDP has an obvious downward trend, due to the fact that our observation window spans across the global financial crisis, which hit Italy particularly hard starting from the first quarter of 2009. However, the average expenditure in the observed supermarket has not been affected at all. Also the number of items has not been affected. If we calculate the corresponding correlations, we notice a negative relationship which, however, fails to pass a stringent null hypothesis test ($p > 0.01$).

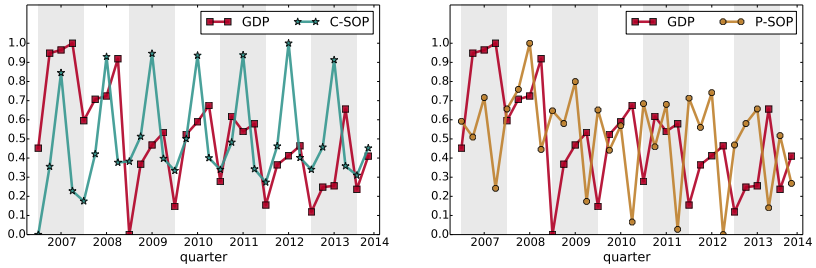

Fig. 6: The relation between GDP and IQM customer (left) and product (right) sophistication.

Turning to our sophistication measure, Figure 6 depicts the relation between GDP and our complex measures of sophistication. On the left we have the measure of customer sophistication we discussed so far. We can see that the alignment is indeed not perfect. However, averaging out the seasonal fluctuation, customer sophistication captures the overall downward trend of GDP. The financial crisis effect was not only a macroeconomic problem, it also affected the sophistication of the satisfiable needs of the population. Note that, again, we have a negative correlation. This means that, as GDP shrinks, customers become more sophisticated. This is because the needs that once were classified as basic are not basic

any more, hence the rise in sophistication of the population. Differently from before, the correlation is actually statistically significant ($p < 0.01$).

We also report on the left the companion sophistication measure: since we can define the customer sophistication as the average sophistication of the products they purchase, we can also define a product sophistication as the average sophistication of the customers purchasing them. Figure 6 (right) shows the reason why we do not focus on product sophistication: the overall trend for product sophistication tends to be the opposite of the customer sophistication. This anti-correlation seems to imply that, as the customers struggle in satisfying their needs, the once top-sophisticated products are not purchased any more, lowering the overall product sophistication index. However, this is only one of many possible interpretations and we need further investigation in future works.

Measure \ Shift	-3	-2	-1	0	1	2
EXP	-0.29302	-0.49830	-0.53078*	0.23976	-0.27619	-0.37073
PUR	-0.27091	-0.49836*	-0.53046**	0.18638	-0.30909	-0.32432
C-DEG	0.24624	0.39808	-0.55479*	0.13727	0.08191	0.36001
P-DEG	-0.12409	-0.26289	-0.57657**	0.30255	-0.22198	-0.28325
C-SOP	-0.32728	**-0.67007*****	0.23261	0.09251	-0.15844	-0.58773**
P-SOP	-0.02675	-0.12916	0.60974**	-0.18587	0.15342	-0.03843

$^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$

Table 3: The correlations of all the used measures with GDP at different shift values. We highlight the statistically significant correlations.

We sum up the correlation tests performed in Table 3. In the Table, we report the correlation values for all variables. We test different shift values, where the GDP timeline is shifted of a given number of quarters with respect to the tested measure. When shift = -1, it means that we align the GDP with the previous quarter of the measure (e.g. GDP Q4-08 aligned with measure's Q3-08).

We also report the significance levels of all correlations. Note that all p-values are being corrected for the multiple hypothesis test. When considering several hypotheses, as we are doing here, the problem of multiplicity arises: the more hypotheses we check, the higher the probability of a false positive. To correct for this issue, we apply a Holm-Bonferroni correction. The Holm-Bonferroni method is an approach that controls the family-wise error rate (the probability of witnessing one or more false positive) by adjusting the rejection criteria of each of the individual hypotheses [15]. Once we adjust the p-values, we obtain the significance levels reported in the table. Only one correlation passes the Holm-Bonferroni test for significance at $p < 0.01$ and it is exactly the one involving the customer sophistication with shift equal to -2. This correlation is highlighted in bold in Table 3, and it represents the main result of the paper.

Note that in the table we also report the correlation values using the IQM for the customer and product degree measures, of which we have not shown the

timelines, due to space constraints. We include them because, as we discussed previously, our sophistication measures are corrected degree measures. If the degree measures were able to capture the same correlation with GDP there would be no need for our more complex measures. Since the degree measures do not pass the Holm-Bonferroni test we can conclude that the sophistication measures are necessary to achieve our results.

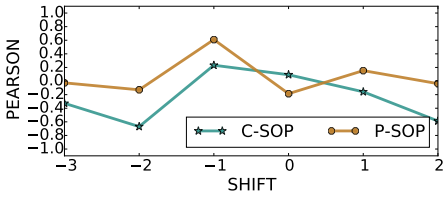

Fig. 7: The correlation between average customer sophistication and GDP with different shifting values.

We finally provide a visual representation of the customer and product sophistication correlations with GDP at different shift levels in Figure 7. The figure highlights the different time frames in which the two measures show their predictive power over GDP. The customer sophistication has its peak at shift equal to -2. The cyclic nature of the data implies also a strong, albeit not significant, correlation when the shift is equal to 2. Instead, the product sophistication obtains its highest correlation with GDP with shift equal to -1. This might still be useful in some cases, as the GDP for a quarter is usually released by the statistical office with some weeks of delay.

6 Conclusion

In this paper we tackled the problem of having a fast and reliable test for estimating the well-being of a population. Traditionally, this is achieved with many measures, and one of the most used is the Gross Domestic Product, or GDP, which roughly indicates the average prosperity of the citizens of a country. GDP is affected by several issues, and here we tackle two of them: it is a hard measure to quantify rapidly and it does not take into account all the non-tangible aspects of well-being, e.g. the satisfied needs of a population. By using retail information, we are able to estimate the overall sophistication of the needs satisfied by a population. This is achieved by constructing and analyzing a customer-product bipartite network. In the paper we show that our customer sophistication measure is a promising predictor of the future GDP value, anticipating it by six months. It is also a measure less linked with the amount of richness around a person, and it focuses more on the needs this person is able to satisfy.

This paper opens the way for several future research tracks. Firstly, in the paper we were unable to define a proper seasonal adjustment for our sophistication measure. The seasonality of the measure is evident, but it is not trivial how to deal with it. A longer observation period and a new seasonal adjustment measure is needed and our results show that this is an worthwhile research track. Secondly, we showed that there is an interesting anti-correlation between the aggregated sophistication measures calculated for customers and products. This seems to imply that, in harsh economic times, needs that once were basic become sophisticated (increasing the overall customer sophistication) and needs that were sophisticated are likely to be dropped (decreasing the overall product sophistication). More research is needed to fully understand this dynamic. Finally, in this paper we made use of a quarterly aggregation to build our bipartite networks. We made this choice because the quarterly aggregation is the most fine-grained one we can obtain for GDP estimations. However, now that we showed the correlation, we might investigate if the quarterly aggregation is the most appropriate for our analysis. If we can obtain comparable results with a lower level of aggregation (say monthly or weekly) our well-being estimation can come closer to be calculated almost in real-time.

Acknowledgements

We gratefully thank Luigi Vetturini for the preliminary analysis that made this paper possible. We thank the supermarket company Coop and Walter Fabbri for sharing the data with us and allowing us to analyse and to publish the results. This work is partially supported by the European Community's H2020 Program under the funding scheme FETPROACT-1-2014: 641191 CIMPLEX, and INFRAIA-1-2014-2015: 654024 SoBigData.

References

1. B. Balassa. Trade liberalization and 'revealed' comparative advantage. *Manchester School*, 33:99–123, 1965.
2. G. Caldarelli, M. Cristelli, A. Gabrielli, L. Pietronero, A. Scala, and A. Tacchella. A network analysis of countries export flows: Firm grounds for the building blocks of the economy. *PLoS ONE*, 7(10):e47278, 10 2012.
3. Sanjay Chawla. Feature selection, association rules network and theory building. *Journal of Machine Learning Research - Proceedings Track*, 10:14–21, 2010.
4. Hyunyoung Choi and Hal Varian. Predicting the present with google trends. *Economic Record*, 88(s1):2–9, 2012.
5. Robert Costanza, Ida Kubiszewski, Enrico Giovannini, Hunter Lovins, Jacqueline McGlade, Kate E Pickett, K Vala Ragnarsdóttir, Debra Roberts, Roberto De Vogli, and Richard Wilkinson. Time to leave gdp behind. 2014.
6. Matthieu Cristelli, Andrea Gabrielli, Andrea Tacchella, Guido Caldarelli, and Luciano Pietronero. Measuring the intangibles: A metrics for the economic complexity of countries and products. *PloS one*, 8(8):e70726, 2013.

7. Claudia Foroni and Massimiliano Marcellino. A comparison of mixed frequency approaches for nowcasting euro area macroeconomic aggregates. *International Journal of Forecasting*, 30(3):554–568, 2014.
8. J W Galbraith and G Tkacz. Nowcasting gdp with electronic payments data. 2015.
9. Liqiang Geng and Howard J. Hamilton. Interestingness measures for data mining: A survey. *ACM Comput. Surv.*, 38(3):9+, 2006.
10. Domenico Giannone, Lucrezia Reichlin, and David Small. Nowcasting: The real-time informational content of macroeconomic data. *Journal of Monetary Economics*, 55(4):665–676, 2008.
11. Riccardo Guidotti. Mobility ranking-human mobility analysis using ranking measures. 2013.
12. Riccardo Guidotti, Michele Coscia, Dino Pedreschi, and Diego Pennacchioli. Behavioral entropy and profitability in retail. *DSAA*, 2015.
13. Ricardo Hausmann, Cesar Hidalgo, Sebastián Bustos, Michele Coscia, Sarah Chung, Juan Jimenez, Alexander Simoes, and Muhammed Yildirim. The atlas of economic complexity. *Boston. USA*, 2011.
14. Dirk Helbing and Stefano Balietti. How to create an innovation accelerator. *The European Physical Journal Special Topics*, 195(1):101–136, 2011.
15. Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70, 1979.
16. Philip A Lawn. A theoretical foundation to support the index of sustainable economic welfare (isew), genuine progress indicator (gpi), and other related indexes. *Ecological Economics*, 44(1):105–118, 2003.
17. Philip A Lawn. An assessment of the valuation methods used to calculate the index of sustainable economic welfare (isew), genuine progress indicator (gpi), and sustainable net benefit index (snbi). *Environment, Development and Sustainability*, 7(2):185–208, 2005.
18. David M Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani. The parable of google flu: Traps in big data analysis. 2014.
19. Chrysa Leventi, Jekaterina Navicke, Olga Rastrigina, and Holly Sutherland. Nowcasting the income distribution in europe. 2014.
20. Alejandro Llorente, Manuel Cebrian, Esteban Moro, et al. Social media fingerprints of unemployment. *arXiv preprint arXiv:1411.3140*, 2014.
21. Brian C Monsell. Update on the development of x-13arima-seats. In *Proceedings of the Joint Statistical Meetings: American Statistical Association*, 2009.
22. Jekaterina Navicke, Olga Rastrigina, and Holly Sutherland. Nowcasting indicators of poverty risk in the european union: a microsimulation approach. *Social Indicators Research*, 119(1):101–119, 2014.
23. Diego Pennacchioli, Michele Coscia, Salvatore Rinzivillo, Fosca Giannotti, and Dino Pedreschi. The retail market as a complex system. *EPJ Data Science*, 3(1):1–27, 2014.
24. Diego Pennacchioli, Michele Coscia, Salvatore Rinzivillo, Dino Pedreschi, and Fosca Giannotti. Explaining the product range effect in purchase data. In *Big Data, 2013 IEEE International Conference on*, pages 648–656. IEEE, 2013.
25. Jameson L Toole, Yu-Ru Lin, Erich Muehlegger, Daniel Shoag, Marta C Gonzalez, and David Lazer. Tracking employment shocks using mobile phone data. *arXiv preprint arXiv:1505.06791*, 2015.
26. N Wilson, K Mason, M Tobias, M Peacey, QS Huang, and M Baker. Interpreting google flu trends data for pandemic h1n1 influenza: the new zealand experience. *Euro surveillance: bulletin européen sur les maladies transmissibles= European communicable disease bulletin*, 14(44):429–433, 2008.