UNIVERSITÀ DI PISA

Facoltà di Scienze Matematiche Fisiche e Naturali
Facoltà di Economia
Corso di Laurea Specialistica in
Informatica per l'Economia e per l'Azienda

Tesi di laurea

Nowcasting Well-Being With Retail Market Data

Candidato:
Luigi Vetturini

Relatori:
Prof. Dino Pedreschi

Dott. Diego Pennacchioli

Dr. Michele Coscia

Controrelatore:
Prof. Roberto Bruni

Anno Accademico 2012/2013

Abstract

The main pourpose of this work is to introduce a novel approach to estimate the degree of human wellness in a flexible and immediate way. Data, provided by Unicoop Tirreno, contains sales informations between 2007 and 2012 and are modeled in a bipartite graph customer- product. We assign to every node of the graph a value called sophistication index, able to show how much a product can be basic or sophisticated and, on the other hand, the propensity of a customer to buy basic or even sophisticated products. We evaluated the temporal evolution of the sophistication index for both customers and products. The performance of the new indicator has been evaluated comparing these trends with the trend of the GDP (Gross Domestic Product) using the statistical method of the Pearson Correlation.

Riassunto

Lo scopo principale del lavoro è di proporre un approccio innovativo per stimare il grado di benessere della popolazione in maniera flessibile e immediata. I dati, forniti dall'Unicoop Tirreno, contengono informazioni sugli acquisti dal 2007 al 2012 e vengono modellati attraverso un grafo bipartito cliente-prodotto. Ad ogni nodo del grafo viene assegnato un valore, chiamato indice di sofisticazione, in grado da un lato di mostrare quanto un prodotto sia base o sofisticato e, dall'altro, quanto un cliente tenda a comprare prodotti base o prodotti sofisticati.

Abbiamo valutato l'evoluzione temporale della complessità economica sia dei clienti che dei prodotti aggregando su entrambi con varie misure statistiche.

L'efficacia dell'indicatore stata valutata paragonandolo con l'andamento del PIL (Prodotto Interno Lordo), attraverso il metodo statistico dell' indice di correlazione di Pearson.

Contents

Introduction

In the modern world politicians and statisticians increased their attention in order to measure population well-being, but the main problem of this measurement is that the known indexes are subject to a lot of criticism. They require a lot of resources, approximations and time to be calculated, so there's always the need of a new way to calculate this wellness. We want to introduce a new index capable to avoid these well-known problems, simple and fast to calculate. This is possible thanks to the wider availability of data in recent years and to the studies for obtaining valuable knowledge from very large databases. These sets of data are called *Big Data*, due to the quantity and complexity of the informations they carry.

Examples of *Big Data* are informations coming from users of social networks, customers of retail companies and query logs of search engines.

It's easy to understand the importance, for big companies, to store and study data, in order to react efficiently to changes in an increasingly dynamic world. The problem is that the *Big Data* become awkward to be managed by standard statistical software, therefore the studies of data scientists has been focused on new sophisticated techniques and algorithms to retrieve informations from these datasets in a tolerable elapsed time. In this thesis we consider sales data provided by Unicoop Tirreno, one of the largest Italian retail distribution Company, to calculate a new economic index capable to show the measure of wellness of a population in different aggregation levels: spatial, temporal, demographic and so on. This analisys cannot be afforded using simple aggregations of values just like quantity of products purchased or total amount of money spent in a period, but we need a more complex index in order to summarize all hidden informations in these data. After a brief survery of the current indexes to measure the well-being, we chose the GDP as a baseline. We, then, decided to use the sophistication index to try to approximate the above measure. Since the mechanism to calculate the GDP requires a large effort in terms of resources and time, we want to try to predict the official releases of this indicator with the index that we chose, in real time (this kind of prediction is called nowcasting in literature).

Organization of the Thesis

In this thesis we want to show how it's possible to significantly approximate well-being indexes. Thesis is organized in the following manner: in chapter 1 will be shown the existing economic indexes to measure population well-being.

Chapter 2 describes the concept of nowcasting and some applications in micro- and macro-economic environment.

In chapter 3 we will introduce the Economic Complexity method to calculate sophistication index, providing some example of application taken from the existing literature.

Chapter 4 is dedicated to the description of the dataset and the data preparation for experiments.

Chapter 5 shows how we calculated the sophistication index, starting from sales data, modeling these data with bipartite graphs and studying the temporal evolution of the calculated index. In this chapter we will validate results using statistical methods.

Chapter 6 contains conclusions and a summary of possible applications of Economic Complexity method.

Chapter 1

Measuring Human Wellness

In this chapter we will describe the most known ways to describe the well-being of a Nation. The need of monitoring economic wellness of Nations, in a macroeconomic enviroment, has always been a central part in economic studies, but the first universally accepted index of economic condition of a Nation, the Gross Domestic Product (GDP), was created only after the *Wall Strett crash* (1929 October 29).
The GDP was subject of a hard criticism so, in last years, a lot of alternatives were processed.

1.1 GDP

GDP was first developed by Simon Kuznets for a US Congress report in 1934, as requested by Franklin Delano Roosevelt. GDP is one of the most comprehensive and closely watched economic statistics: it is used by the White House and Congress to prepare the Federal budget, by the Federal Reserve to formulate monetary policy, by Wall Street as an indicator of economic activity, and by the business community to prepare forecasts of economic perfor mance that provide the basis for production, investment, and employment planning [1].

1.1.1 Measuring GDP

There are three approaches to measuring GDP

1. **Income Approach.** Sum total of incomes of individuals living in a country during 1 year. GPD, calculated in this way, is sometimes called Gross Domestic Income (GDI) and should provide same amount as the expenditures approach, but measurement errors will make the two values of GDP and GDI slightly different.
   There are five categories of incomes:

   (a) Wages: The official measure of wages earned by the household sector for supplying labor services. Compensation of employees is far away the largest of the five factor payments, typically running about 55 percent of gross domestic product. It includes standard wages and salaries paid directly to employees, as well as fringe benefits paid on behalf of employees to third parties.

   (b) Corporate profits: The total accounting profits received by corporations, which is the official measure of profit earned by the household sector for supplying entrepreneurship services through corporations. Corporate profits are the second largest factor payment category, usually coming in around 15 to 20 percent of Gross Domestic Product.

   (c) Net Interest: The official measure of interest earned by the household sector for supplying capital services. Net interest is usually less than 8 percent of Gross Domestic Product, typically in the 5 to 7 percent range. It is revenue generated from borrowed funds but is considered payment for the productive services of capital.

(d) Proprietors' Income: Is the money earned by business owners and includes payments to all factors of production-labor, capital, land, and entrepreneurship. Proprietors' income is usually less than 8 percent of Gross Domestic Product, typically falling in the 6 to 9 percent range.

(e) Rental Income of Persons: The official measure of rent earned by the household sector for supplying land and related services. Rental income of persons is typically the smallest of the five factor payment categories, usually less than 4 percent of Gross Domestic Product. It includes payments for the use of land, natural resources, and capital goods attached to the land.

Summing these five factors we will have the *National Income(NI)* and three adjustments are needed to get GDP. First of all, government receipts are not part of the GDP, because income tax receipts include money that is part of the incomes of other segments of echonomy and they are already being counted elsewhere. However, some indirect buisiness taxes need to be added to te equation to get the *Net National Product(NNP)*.

$$NNP = NI + Indirect\ taxes$$

GDP calculated using expenditures approach includes gross private domestic investment. Not all of this amount is received as income, but some of it is used to replace worn-out or damaged equipment. This replacement value is called *Capital Consumption Allowance(Depreciation)* and, in order to balance income and expenditures, needs to be added to income. Adding the depreciaton to NNP we will get a number called *Gross National Product(GNP)*.

$$GNP = NNP + Capital\ Consumption\ Allowance(Depreciation)$$

Last adjustment to calculate GDP is made taking income received by citizens outside the nation's borders, and subtracting income received by foreigners within the nation's borders. This adjustment is called *Net Factor Income From Abroad* and has to be subtracted to the GNP

$$GDP = GNP \text{ - } Net\ Factor\ Income\ From\ Abroad$$

2. **Exspenditures Approach.** Sum all expenditure incurred by individuals during 1 year. GDP is a sum of Consumption (C), Investment (I), Government Spending (G) and Net Exports.

   $$GDP = C + I + G + NX$$

   Net Exports (NX) is equal to total exports minus total imports. Exports are added into GDP because they represent goods and services that are produced within the economy but are not part of domestic expenditures. Imports are subtracted because they represent spending on goods and services that were not produced within the economy. We can rewrite the GDP formula in this way:

   $$GDP = C + I + G + (X\text{-}M)$$

   where X stands for total exports and M stands for total imports.

3. **Value Added Approach.** GDP is the market value of all final goods and services calculated during 1 year. This method requires three stages of analysis. First gross value of output from all sectors is estimated. Then, intermediate consumption such as cost of materials, supplies and services used in production final output is derived. Then gross output is reduced by intermediate consumption to develop net production.

1.1.2 Limitations of GDP

The first criticism over GDP was moved by its own creator, Simon Kuznets, who refuses the use of this index as wellness measurement. Everything that can be sold and has a monetary value will increase the GDP. Another criticism come from Robert Kennedy, that stated: *"It counts special locks for our doors and the jails for the people who break them. It counts the destruction of the redwood and the loss of our natural wonder in chaotic sprawl. It counts napalm and counts nuclear warheads and armored cars for the police to fight the riots in our cities. It counts Whitman's rifle and Speck's knife, and the television programs which glorify violence in order to sell toys to our children ".* By these words we can understand that a growth of GDP doesn't necessarily means increasing of individual and collective well-being. Well-being losses are not measured anywhere in GDP, the destruction of the Amazon Forest is an activity that makes the world GDP increase. The resulting loss of natural capital, its effects on the climate, biodiversity, and the long-term needs of future generations are not measured anywhere. In other words, the GDP does not deduct the losses of natural capital, but makes additions to account for its organized destruction. On the other side, many activities that contribute to the well-being are not counted, for example volunteer work and household work.

GDP measures the amount of goods produced, but ignores satisfaction resulting from the consumption of such goods. In United States, since 1980, the average annual working hours has risen five hours a year, as opposed to what has happened to almost all European countries, in this way, forcing people to work more, we will increase GDP, but we will have less free time and well-being derived by it.

Last - and central critic in this thesis pourpose - is that GDP ignore distribution of richness and poverty. We don't know, by looking the average GDP, how this income is shared among the people. A growth of GDP may come togeter with an increase or reduction in social inequalities.

1.2 Alternative Indexes of Economic Progress

The criticism over GDP didn't prevent politicians to use GDP as the main instrument to measure economic growth. On the other side, some economists have developed new indexes for richness and wellness measurement. The main examples are:

- Index of Sustainable Economic Welfare (ISEW)
- Genuine progress indicator (GPI)
- Human development index (HDI)

1.2.1 Index of Sustainable Economic Welfare (ISEW)

The ISEW is one of the most advanced attempts to create an indicator of economic welfare. It was originally developed in 1989 by Herman Daly and John B. Cobb, but later they went on to add several other costs to the definition of ISEW.This later work resulted in another macroeconomic indicator, Genuine Progress Indicator (GPI), that will be presented below.

The Index of Sustainable Economic Welfare (ISEW) is roughly defined by the following formula.

ISEW = personal consumption
+ public non-defensive expenditures
- private defensive expenditures
+ capital formation
+ services from domestic labour
- costs of environmental degradation
- depreciation of natural capital.

- Personal consumption: includes all consumption goods consumed by private households. Personal consumption expenditure is weighted according to changes in the distribution of income, otherwise it will inaccurately reflect its true contribution to a nations economic welfare.

- Public non-defensive expenditures: Daly and Cobb included only that fraction of public expenditures on health and education which they believed to represent non-defensive consumption expenditures. On this basis, ISEW includes one half of all medical expenditures, the excluded half being assumed to be defensive expenditures, and one half of all higher education expenditures, which are assumed to represent pure consumption.

- Private defensive expenditures: is the sum of the costs of pollution control, costs of car accidents, costs of noise pollution and costs of commuting. In some cases, a certain percentage of private health expenditure is assumed to constitute a form of defensive expenditure.

- Capital formation: e.g. we can define it as equipment to be used in the future.

- Services from domestic labour: domestic labour for cleaning, cooking and childminding, for example, contributes directly to economic welfare, even if it does not involve money.

- Costs of environmental degradation: includes costs of water and air pollution, costs of climate change and costs of ozone deplation.

- Depreciation of natural capital: is the sum of the cost of loss of farmlands, in terms of quantity and quality, and the cost of depletion of non-renewable resources (oil, natural gas and so on) valued at a replacement cost with renewable substitutes estimate plus an escalation factor.

1.2.2 Genuine Progress Indicator (GPI)

GPI was developed as an evolution of ISEW,

The economists Herman Daly, John B. Cobb and Philip Lawn have asserted that a country's growth, in terms of a major production of goods or services, has both costs and benefits, and not just the benefits that contribute to GDP. In some situations, the expanded production can damage the health, culture, and welfare of people.

Lawn proposed a model that includes the following potential harmful effects in a country's growth[2]:

- Cost of resource depletion
- Cost of crime
- Cost of ozone depletion
- Cost of family breakdown
- Cost of air, water and noise pollution
- Loss of farmlands
- Loss of wetlands

GPI takes into account these problems by incorporating sustainability, for example an activity that pay attemption to air pollution will score a better GPI than a similar activity without this pollution-care. Compareing GPI to the GDP, we can assert that GPI is the GDP minus the environmental and social costs, infact, in the same example of air pollution, GDP gains when pollution is created as side effect of a valuable activity and gains when pollution is cleaned up, whereas GPI assign a negative weight to pollution creation. The main problem is that this weights are difficult to extimate.

The formula of GPI, in simplified form, is:

$$GPI = A + B - C - D + I$$

where:

- A= income weighed private consumpion
- B = value of non-marke services generating welfare
- C = privae difensive cos of natural deterioration
- D = cost of deterioraion of nature and natural resources
- I = increase in capital stock and balance of internaional trade

The single components of the GPI:

- + Personal consumption weighted by income distribution index
- + Value of household work and parenting
- + Value of higher education
- + Value of volunteer work
- + Services of consumer durables
- + Services of highways and streets
- - Cost of crime
- - Loss of leisure time
- - Cost of unemployment
- - Cost of consumer durables
- - Costs of commuting
- - Cost of household pollution abatement
- - Costs of transport accidents

- - Costs of industrial accidents
- - Cost of water pollution
- - Costs of air pollution
- - Costs of noise polluion
- - Costs of loss of wetlands
- - Costs of loss of farmlands
- -/+ Loss of forest area and damage from logging roads
- - Depletion of nonrenewable energy resources
- - Carbon dioxide emissions damage
- - Costs of ozone depletion
- +/- Net capital investment
- +/- Net foreign borrowing

1.2.3 Human Development Index (HDI)

The Human Development Index[3] is a summary of human development around the world and implies whether a country is developed, still developing, or underdeveloped based on factors such as life expectancy, education, literacy, gross domestic product per capita.

It was created by the Pakistani economis Mahbub ul Haq and the Indian economist Amartya Sen in 1990 wih the explicit purpose “to shift the focus of development economics from national income accounting to people-centered policies.”

We can summarize the HDI as the average of three normalized variables: income, longevity and education. For each variable, the normalization is:

$$X = \frac{\text{actual value} - \text{minimum value}}{\text{maximum value} - \text{minimum value}}$$

In formulas, HDI is the geometric mean of three index:*Life Expectancy Index (LEI)*, *Education Index (EI)* and *Income Index (II)*.

- LEI

$$LEI = \frac{LE - 20}{83.6 - 20}$$

  LE = life expectancy at birth
  Minimum value for life expectancy is fixed at 20 years and maximum is kept at 83.6 years.

- EI: Education Index is composite of two indexes

$$EI = \frac{\sqrt{MYSI * EYSI} - 0}{0.971 - 0}$$

  Mean Years of schooling Index (MYSI) and *Expeced Years of Schooling Index (EYSI)* are calculated by the following formulas:

$$MYSI = \frac{MYS - 0}{13.3 - 0}$$

  MYS = mean years of schooling (years that a 25 years old person or older has spent in schools).
  The low value was fixed at 0 and the maximum value for mean years of schooling is fixed at 13.3.

$$EYSI = \frac{EYS - 0}{18.0 - 0}$$

  EYS = expeced years of schooling (years hat a 5 years old child will *spend with his education in his whole life).*

Low value for expected years of schooling is fixed at 0 and high value is fixed at 18.0.

- II

$$II = \frac{ln(GNIpc) - ln(100)}{ln(87,478) - ln(100)}$$

GNIpc = gross national income at purchasing power parity per capita

Minimum income is set as \$100 and maximum income is set as \$87,478.

Finally, we can calculate the HDI:

$$HDI = \sqrt[3]{LEI * EI * II}.$$

Reflecting inequality in each dimension of the HDI, has been introduced the Inequality-adjusted HDI (IHDI), a measure of the level of human development of people in a society that accounts for inequality. Under perfect equality the IHDI is equal to the HDI, but falls below the HDI when inequality rises. The IHDI accounts for inequality in HDI dimensions by discounting each dimensions average value according to its level of inequality measured by the Atkinson index.

1.3 Criticism Over Described Methods

We already discussed about the criticism over GDP in section 1.1.2, but the alternative methods aren't exent from criticism. The main criticism over ISEW is that financial costs have to be assigned to non-financial impacts such as climate change and ozone depletion and this use of such nonstatistical judgements invalidates the utility of ISEW. Similar to the ISEW, GPI has drawn criticism based on the inconsistent and somewhat arbitrary list of adjustment items, as well the monetary valuation methods used to measure aspects of well-being outside the market[4]. Responses to these criticisms have defended GPI methods and provided a more solid theoretical foundation for the GPI and related indicators[5]. An additional point is that many of the choices involved in GDP accounting are equally arbitrary but, as with GPI, are justified according to the goals of the measure being constructed. It is also important to point out the height of inconsistency and arbitrariness to use GDP (a measure of activity or income) as a measure of welfare, something for which it was never intended[6]. The HDI tries to overcome this problems with less subjective assessments, but still subjective. The common problem, shared by all the described methods, is purely related to the use of resources in terms of time and persons working on it, for example the value of the GDP is subject to several adjustment for some years before the real value can be calculated. Despite the criticism over the GDP, in this thesis we decided to use it as a reference point, as it is an index with a few approximations, using less abstract values and also it's easy to retrieve the data.

In this chapter we described the existing indexes to measure human wellness, we illustrated the methods to calculate them and the limitation of these methods. In next chapter we will introduce the concept of *nowcasting*, a way to *predict the present*.

Chapter 2

Nowcasting: using *Big Data* to predict the present

As we presented above, one of the criticism over the existing indicators of the level of the economic activity is that they are normally relased with several weeks of delay and often they keep beeing revised and adjusted for months after their publication.

This problem drove economists to study a new kind of approach to more efficently forecast the economic activity. In this chapter we will explain the concept of "nowcasting" and we will present two examples of application: one of the most relevant about nowcasting ([7]) and a short example of the use of nowcast in relation with GDP ([8]).

2.1 Nowcasting Using Google Trends

In [7] the authors studied "Google Trends", a real time daily and weekly index of the volume of queries that users enter into Google and they clame that this index is not able to predict the future, but is able to "predict the present". For example, the volume of queries on automobile sales during the second week in June may be helpful in predicting the June auto sales report which is released several weeks later in July. They found that queries can be useful leading indicators for subsequent consumer purchases in situations where consumers start planning purchases significantly in advance with respect to their actual purchase decision.

The authors call this "contamporaneus forecast" with the term *"nowcasting"*, created from unification of the words "now" and "casting" and first used in meteorology for predicting the weather of the next hours by using recent data from satellites and weather stations.

Google Trends provides a time series index of the volume of queries users enter into Google in a geographic area. The index is calculated dividing the total volume of the search term within a particular geographic region for the number of the queries in that region during the examinated period. This number is normalized between 0 and 100.

The query are "broad matched": for exampe, the query *used automobiles* are counted in the calcolation of the query index for automobile.

Google classifies search queris into about 30 categories at the top level and about 250 categories at the second level using a natural language classification engine. For example, the query [car tire] would be assigned to category *Vehicle Tires* which is a subcategory of *Auto Parts* which is a subcategory of *Automotive.* Some query could be assigned to different categories (for example [apple] can be assigned to *Computers & Electronics*, *Food & Drink*, and *Entertainment*) so this assignment is probabilistc.

They use the nowcasting application of Google Trends in four examples: motor vehicles and parts sales, initial claims for unemployment benefits, travel data to predict visits to a particular destination and the Roy Morgan Consumer Confidence Index for Australia.

2.1.1 Mothor Vehicles and Parts

The first example provided by authors is about “Motor Vehicles and Parts Dealers” series from the U.S. Census Bureau “Advance Monthly Sales for Retail and Food Services” report. The index summarizes results from a survey sent to motor vehicle and parts dealers that asks about current sales. The data is available in both seasonally adjusted and unadjusted form and they used the unadjusted data. The preliminary index is released 2 weeks after the end of each month. Let y_t be the log of the observation at time t, they fist estimate a simple baseline seasonal AR-1 model $y_t = b_1 y_{t-1} + b_{12} y_{t-12} + e_t$ for the period 2004-01-01 to 2011-07-01. Google Trends variables can improve out-of-sample forecasting? To check it they used a rolling window forecast where they estimated the model using the data for periods k through $t-1$ and then forecast $y-t$ using y_{t-1}, y_{t-12}, and the contemporaneous values of the Trends variables as predicors. They use $k = 17$ to have a reasonable number of observations for the regression.

The mean absolute error (MAE) of $log(y_t)$ using the baseline seasonal AR-1 model is 6.34% while MAE using the Trends data is 5.66%, with an improvement of 10.6%.

2.1.2 Initial claims for unemployment benefits

Each Thursday morning the US Department of Labor releases a report describing the number of people who filed for unemployment benefits in the previous week.

When someone becomes unemployed, probably they will search something like *file for unemployment*, *unemployment office*, *jobs* and so on. Google Trends classifies these queries into two categories: *Local/Jobs* and *Society/Social Services/Welfare & Unemployment.* In this sample the authors use the seasonally adjusted initial claims data, since that is the number used by most economic forecasters, so they seasonally adjusted the Trends data as well. Their baseline regression is a simple AR-1 model on the log of initial claims. The MAE, using the baseline forecasts, is 3.37%, while goes to 3.68% using the Trends data and that means a 5.95% reduction in fit. To explain it they examine the series a bit more closely.

It is well-known that it is difficul to identify "turning points" in conomic series. They identify 4 turning points in the sequence and there is a reduction in MAE at all turning points, especially in the first two, so the Google Trends data seems to help in identifying at least two turning points.

2.1.3 Travel

The internet is commonly used for travel planning which suggests that Google Trends data about destinations may be useful in predicting visits to that destination. They analyze data provided by the Hong Kong Tourism Board, that publishes monthly visitor arrival statistics, including monthly visitor arrival summary by country of residence. The authors use visitor data from Us, Canada, Great Britain, Germany, France, Italy, Australia, Japan and India.

Hong Kong is also one of the subcategories in Vacation Destinations in Google Trends and is possible to examine the query index for this category by country of origin.

For this example they use not seasonally adjusted data. They use the average query index in the first two weekly observations of the month to predict the total monthly visitors, with 6 weeks of forecas, cause the data is released with a one-month lag. Setting y_t as the number of visitors from a given country in month t and x_t as the average Google Trends index for *Vacation Destinations/Hong Kong* for the first two weeks of the month, the seasonal AR-1 model that they use is: $y_t = b_1 y_{t-1} + b_{12} y_{t-12} + b_0 y_t + e_t$.

They estimate this model for each country and compare the actual with the fitted results. In this case the authors used in-sample fits The fits are pretty good, with the exception of Japan and the average R^2, without Japan, is 73.3%.

2.1.4 Consumer Confidence

In this final example, the authors examine the Roy Morgan Consumer Confidence Index for Australia. It's not easy to say what categories are useful in predicting the series, so they used a Bayesian method known as "spike and slab" regression that produces a posterior probability that a variable enters a regression. This method assignes high posterior probabilities to the Google Trends Categories *Crime & Justice*, *Trucks & SUVs* and *Alternative Vehicles*. The last two are not surprising, cause they are correlated with the price of gasoline, wich is known to impact consumer confidence, but there is no explanation for the first predictor.

The Trends predictors reduce MAE of the simple AR-1 model by about 12.7% for in-sample forecasts. One-step-ahead MAE goes from 3.63% to 3.29%

2.2 Nowcasting GDP

Now the idea is to find some way to nowcast the well-being indexes described in 1. In [8] the authors apply che concept of nowcasting for evaluating the marginal impact that intra-monthly data releases have on current-quarter forecasts (nowcasts) of real GDP growth. The idea is to use small models to bridge the information contained in one or a few key variables with monthly aggregation with the quarterly growth rate of GDP, which is released after data. For this pourpose they provide a statistical framework able to combine a large amount of data that, besides nowcasting GDP, can be used to evaluate the impact of each new data release on the nowcast and its accuracy. This framework combines three aspects of nowcasting: it uses a large number of data series, it updates the nowcasts and the measures of their accuracy in accordance with the real-time calendar of data releases, and it bridges monthly data releases with the nowcast of quarterly GDP.

The first issue of the authors was to choose the variables to build the model. Since the number of variables in the information set is large, estimating a full model would limit the degrees of freedom and hence the model would perform poorly in forecasting because of the large uncertainty in the parameters' estimation, so they need to reduce the number of the variables. Moreover, in real time, some data series have observations through the current period, whereas for others the most recent observations may be available only for a month or quarter earlier. Consequently, the underlying data sets are unbalanced. To deal with these problems, they tried to adapt a "factor model", a regression method now standard at central banks and other institutions to explain some phenomena using many factors, with a two steps method. First, they extimated the parameters of the model with an "Ordinary least Squares" regression on the data truncating the data set at the date of the least timely release. In the second step, the common factors are extracted by applying the "Kalman smoother" on the entire data set. The Kalman smoother is a technique that uses a series of measurements observed over time, containing noise (random variations) and other inaccuracies, used to compute recursively the expected value of the common factors ([9]).

The model is used to produce nowcasts based on about 200 time series for the US economy typically used by short-term forecasters. By tracking the calendar of data releases throughout each quarter, it produces a nowcast of GDP corresponding to each data release. This sequence of nowcasts is used to evaluate the nowcasts accuracy as the set of variables used in the predictions evolves over time and to assess the real-time marginal impacts that different types of economic information have on the nowcast of GDP.

To examine the performance of the model, they perform two sets of exercises. In the first, they provide an evaluation of the overall performance of the model, in the second, they study the effect of each release during the quarter on the forecast accuracy; i.e., they analyze the evolution of the forecast in relation to the flow of information throughout the quarter.

This work is quite near to the pourposes of this thesys and represents a way to nowcast the well-being, but we want to do it in a simpler manner, and avoiding to dig into that huge multitude of macroeconomic data.

After the description of the existing indexes of well-being, we introduced the concept of nowcasting, reporting some experiments about it. In next chapter we will talk about the methodology of Economic Complexity, that will be used to calculate our index that measures the human well-being and that try to nowcast the existing indexes.

Chapter 3

Economic Complexity

We discussed about existing methods to estimate human well-being and explored all the criticism over them, now we will discuss about a quite new method that can be useful for our purposes and the applications of this method in existing literature. The Economic Complexity method has been applied in two contexts: a microeconomic[10] and a macroeconomic[11].

3.1 Microeconomic Approach

This approach is really similar to the approach of this thesis, studying market basket transactions of UNICoop in order to retrieve useful informations from their shopping activity. The authors create a framework able to explore the bipartite graph G = *(C,P,E)* connecting the customers $c \in C$ to the products $p \in P$ they buy where w_j on the edge $(c_i, p_i, w_i) \in E$ is the number of times customer c_i bought product p_i. This graph was rapresented by its adjacency matrix.

3.1.1 The Adjacency Matrix and the Isocline

In this section we will present how the adjacency matrix was built, the methods they used to build it and the explanation of these methods. Moreover we will talk about the mathematical way they used to describe the resulting matrix.

Starting from the bipartite graph $G =$ *(C,P,E)*, the matrix was obtained placing the customers on the rows, sorted in a descending order on the basis of the sum of the items purchased, and the products on the columns, with the same criteria from left to right. To evaluate how meaningful is a purchase quantity for each product p_i, for each customer c_j, the concept of lift is used. Given a pair of itemsets *(X, Y)*, the lift of the pair is defined as follow:

$$lift(X, Y) = \frac{supp(X, Y)}{supp(Y) * supp(X)},$$

where $supp(I)$ is the relative support of the itemset I. The relative support of itemset I is the number of times all $i \in I$ are purchased together over all the transactions present in the dataset. In this case the itemset X always contains one item (the customer c_j), the itemset Y always contains one element (the product p_i) and the support of $(c_j; p_i)$ is given by the corresponding entry in the matrix. In other words, $supp(c_j, p_i)$ is the relative amount of product p_i bought by customer c_j , $supp(p_i)$ is the relative amount sold of product p_i to all customers and $supp(c_j)$ is the relative amount of products bought by customer c_j. If $lift(c_j, p_i) < 1$ it means that customer c_j purchased the product p_i less than expected, and viceversa, so the value 1 for the lift is a good treshold to discern the meaningfulness of the quantity purchased. The lift M_{cp} matrix is built in this way:

$$M_{cp} = \begin{cases} = 1 & \text{if lift}(c_j, p_i) > 1 \\ = 0 & \text{otherwise} \end{cases}$$

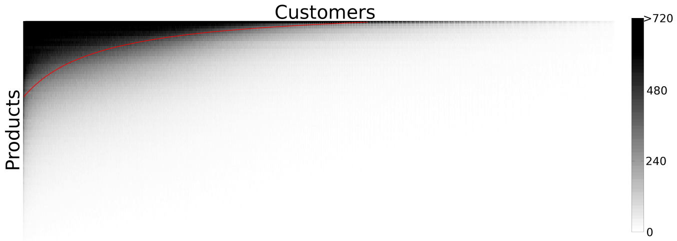

Figure 3.1: The M_{cp} purchase matrix. For layout purposes, the matrix has been transposed, thus we have customers as columns and products as rows. This is a compressed view of the matrix, where each data dot represent a 50x50 square of the original matrix and the gray gradient represents how many 1s are present in that section of the matrix. The red line is the isocline of the matrix.

At a first look, we can see that all the ones in the M_{cp} matrix are placed in the top-left corner. Nevertheless, the matrix has a very particular shape. We can notice, first, that each row is a subset of the above rows, and the same happens for the columns (from left to right). This means that there exists an ordering on the products that leads to a hyerarchical structure based on the products popularity. Secondly, we can notice that the more explicative zones of the matrix are teh top-right corner and the bottom-left corner. On the top-right corner we have those customers that buy very few products and those products are the ones bought by almost all the people. On the other hand, on the bottom-left corner, we have those products that are bought by very few people, and those people buy almost everything. All the above considerations lead to the definition of *sophistication* of a product: the more a product is sophisticated, the less has to be sold, and, moreover, the customers that buy that product need to have bought all the less sophisticated products. A matrix that respect all the above characterstics is defined as *triangular*.

Now there is the need to validate this model and prove that this triangular structure is meaningful. For this goal, the authors use a null model to compare it to the real world data. This null model must hold three features:

1. The purchases are distribuited randomly.

2. Customers must preserve the total amount of their purchases.

3. Each product must preserve its sales volume on the market.

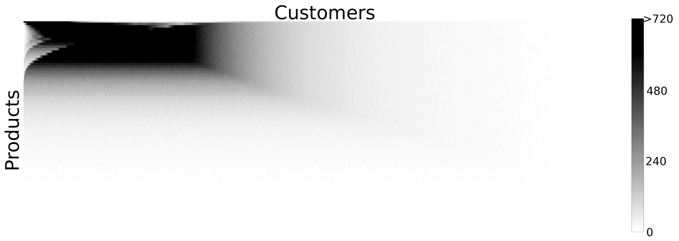

Figure 3.2: One instance of the null purchase matrix. To be consistent with Figure 3.1, also the null matrix has been transposed.

Given these assumptions, they generated a random matrix where the observed sums of rows and columns are preserved using their own algorithm, cause the existing algorithms are not designed to work on large matrices. The obtained null M_{cp} matrix (Figure 3.2) still presents some characteristic of the original M_{cp} matrix, however popular customers/products tend to have randomly distributed lifts and and, while preserving some triangularity, the null model matrix have a tendency to display more ones on the top-left to bottom-right diagonal than the original Mcp matrix. The conclusions are that the null hypothesis explain only part of the observed structure but the original matrix presents some characteristics that cannot be generated randomly. On these characteristics is build the f_* function that characterizes the M_{cp} matrix structure. Taking a look of Figure 3.1, the f_* function is the equation of the line dividing the black area (the one with the high density of ones) from the white area.

There is a strong assumption building this f_* function: the assortment of products bought by any given customer c_j is determined by c_j s volume of purchase, and the population of customers that buy any given product p_i is determined by$p_i$is volume of sales. More precisely, the f_* function relates the rank of a product with the rank of a customer,where rank i(j for customers) stands that a product p_i is the i-th highest sold product (or c_j is the j-th customer with largest volume purchases). The function f_* is a systematic map from the rank j of a customer c_j to the rank $i = f_*(j)$ of a product p_i, such that the assortment of products bought by c_j is $\{p_1; .....; p_i\}$ with high probability. The mapping can be inverted, so we can map the rank i of any product p_i into the rank $j = f^{-1}(i)$ of a customer c_j .

For any customer c_j the authors define $assortment(c_j) = \{p1, ..., p_{f_*(j)}\}$ and for any product p_i, $customer_base(p_i) = \{c_1, ...c_{f_*^{-1}(i)}\}$. The mathematical shape of the f_* map appears to be, from Figure 3.1, anti-monotonic, i.e., $i_1 < i_2$ implies that $f_*(i_1) > f_*(i_2)$, which implies that $assortment(c_2) \subseteq assortment(c_1)$. In other words, if c_1 is customer purchasing more in terms of product quantities than c_2, then it is very likely that c1 buys the same set of products c2 buys, plus something more. In ecology literature, nestdness is defined as a measure to understand how much triangular is a matrix representing the connections between species and ecosystem. The nestedness is calculated by identifying the border dividing the matrix in two areas containing respectively most ones and most zeroes, exactly the role of f_* function. This function is known as *isocline.*

To evaluate if a proposed isocline is good or not the following formula is used:

$$N(M_{cp}, f_*) = \frac{1}{2}\left(\frac{f_l(M_{cp}, 1)}{f_l(M_{cp}, *)} + \frac{f_r(M_{cp}, 0)}{f_r(M_{cp}, *)}\right),$$

where $f_l(M_{cp}, *)$ counts the number of cells at the left of the isocline in M_{cp} where ones are expeted to be found ($f_l(M_{cp}, 1)$ counts the ones) and where $f_r(M_{cp}, *)$ counts the number of cells at the left of the isocline in M_{cp} where zeroes are expeted to be found ($f_r(M_{cp}, 0)$ counts the zeroes).

Now the authors have to estimate where the isocline should pass to maximize the division of ones at the left and zeroes at the right. Considering the matrix as a Cartesian space, for each discrete x axis value (customer) they get an estimate of where the isocline should pass (y axis) summing the ones of the corresponding matrix row $((k_{c,0} = \sum_p M_{cp}(c,p)))$. Then, for each discrete y axis value (product) the estimation of where the isocline should pass (x axis) is obtained summing all ones of the corresponding matrix column $((k_{0,p} = \sum_c M_{cp}(c,p)))$. The two values are averaged and a pair of coordinates is obtained. The coordinates are fitted using a non-linear least squares optimization with the Levenberg-Marquardt algorithm to obtain the best function able to represent the isocline and, therefore, the f_* function. The simple non-rectangular hyperbola is resulted to be the best function for the f_* in the dataset considered.

3.1.2 Calculating the Sophistication

Now the purpose of the work is to quantify sophistication level of the products and of the customers. It's not possible to assert that the more a product is sold, the more basic it is. Another necessary condition is that the set of customers buying the product should include the set of customers with the lowest level of sophistication of their needs. For this reason, the authors of the work need to evaluate at the same time the level of sophistication of both the products and the needs of a customer using the data in the purchase matrix, and recursively correct the one with the other. In the first step they calculeted the sums of the purchase matrix for each customer $(k_{c,0} = \sum_p M_{cp}(c,p))$ and product $(k_{0,p} = \sum_c M_{cp}(c,p))$ and afther they corrected this sums recursively: they calculated the average level of sophistication of the customers by looking at the average sofistication of the products that they buy and then used it to update the sophistication of these products, and so forth. We can summarize the calculation as follows:

$$k_{np} = \frac{1}{k_{0p}} \sum_c M_{cp} k_{c,N-1}$$

Then $k_{c,N-1}$ is inserted into $k_{N,p}$ obtaining:

$$k_{N,p} = \frac{1}{k_{0,p}} \sum_{c} M_{cp} \frac{1}{k_{c,0}} \sum_{p'} M_{cp'} k_{N-2,p'}$$

$$k_{N,p} = \sum_{p'} k_{N-2,p'} \sum_{c} \frac{M_{cp} M_{cp'}}{k_{0,p} k_{c,0}}$$

and rewrite this as:

$$k_{N,p} = \sum_{p'} \widetilde{M}_{p,p'} k_{N-2,p'},$$

where:

$$\widetilde{M}_{p,p'} = \sum_{c} \frac{M_{cp} M_{cp'}}{k_{0,p} k_{c,0}}.$$

$K_{N,p}$ is satisfied when $K_{N,p} = K_{N-2,p}$, and this is equal to a certain costant a. This is the eigenvector associated with the largest eigenvalue, but it's not informative, cause it's composed by the same costant. They look, instead, for the eigenvector associated with the second eigenvalue. This is the eigenvector associated with the variance in the system and can estimate the product sophistication.

This formulation is very sensitive to noise, so they need a strategy to avoid it. They use a three step strategy. First, they calculate the eigenvector on a restricted number of most popular products, then they use the estimate of the sophistication of these products to estimate the sophistication of the entire set of customers (that is the average sophistication of the restricted set of products), and finally they use the estimated sophistication of the customers to have the final sophistication of the entire set of products. The definition of the product sophistication index PS is:

$$PS = \frac{\vec{K} - \mu(\vec{K})}{\sigma(\vec{K})},$$

where $\vec{K}$ is the eigenvetor of $\widetilde{M}_{p,p'}$ associated to the second largest eigenvalue, normalized as described above; $\mu(\vec{K})$ is it's average and $\sigma(\vec{K})$ its standard deviation.

This procedure was applied by the authors to the dataset of retail market data of one of the largest Italian retail distribution company. Considering that the dataset contains more than one million customers and almost 350k items, would generate $\sim$ 370 billions of cells, that is redundant for the purposes of the work, they decided to apply a geographic selection including in the dataset used all the purchases of the customers located in the city of Livorno during the period from 2007 to 2009.

The results were used in three ways: 1) reconstruct the hierarchy of needs of the supermarket customers, 2) provide marketing strategies and, 2) in a similar work ([12]), find a link between sophistication of a need of a customer and the distance he travels to buy the product that can satisfy that need.

3.1.3 Pyramid of Needs

In this section we will show the results applied to build a specific instance of the Maslow's hierarchy of needs using Unicoop data. The Maslow's hierarchy of needs is a theory in psychology propsed by Abrahm Maslow in [13] and is often portrayed in the shape of a pyramid with the largest, most fundamental levels of needs at the bottom and the need for self-actualization at the top 3.3.

The most fundamental and basic four layers of the pyramid contain what Maslow called deficiency needs": esteem, friendship and love, security, and physical needs. If these deficiency needs are not met, with the exception of the most fundamental (physiological) need, there may not be a physical indication, but the individual will feel anxious and tense. Maslow's theory suggests that the most basic level of needs must be met before the individual will strongly desire (or focus motivation upon) the secondary or higher level needs, and this is exactly what happens with sophistication, a customer, before buying high-sophisticated products, buys all the others. The levels

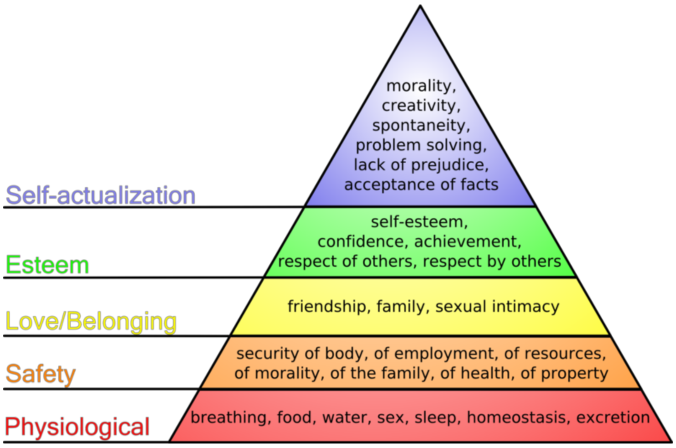
morality,
creativity,
spontaneity,
problem solving,
lack of prejudice,
acceptance of facts
Self-actualization
self-esteem,
confidence, achievement,
respect of others, respect by others
Esteem
friendship, family, sexual intimacy
Love/Belonging
security of body, of employment, of resources,
of morality, of the family, of health, of property
Safety
breathing, food, water, sex, sleep, homeostasis, excretion
Physiological

Figure 3.3: Maslow's Hierarchy of Needs. While the pyramid has become the de facto way to represent the hierarchy, Maslow himself never used a pyramid to describe these levels in any of his writings on the subject.

of the pyramid of Maslow are:

1. Phisiological needs: Physiological needs are the physical requirements for human survival. If these requirements are not met, the human body cannot function properly and will ultimately fail. Physiological needs are thought to be the most important, they should be met first. These needs are metabolic requirements as air, water and food or clothing and shleter to provide protection from the elements.

2. Safety needs: With their physical needs relatively satisfied, the individual's safety needs take precedence and dominate behavior. These needs include: personal security, financial security, health and well-being, safety net against accidents/illnes and their adverse impacts.

3. Love and belonging: After physiological and safety needs are fulfilled, the third level of human needs is interpersonal and involves feelings of belongingness. These needs include relationship such as friendship, intimacy, family. Moreover, according to Maslow, humans need to feel a sense of belonging and acceptance among their social groups, regardless if these groups are large or small.

4. Esteem: All humans have a need to feel respected; this includes the need to have self-esteem and self-respect. Esteem presents the typical human desire to be accepted and valued by others. . Maslow noted two versions of esteem needs: a lower" version and a higher" version. The lower version of esteem is the need for respect from others. This may include a need for status, recognition, fame, prestige, and attention. The higher version manifests itself as the need for self-respect. This higher version takes precedence over the lower.

5. Self-actualization: This level of need refers to what a person's full potential is and the realization of that potential. Maslow describes this level as the desire to accomplish everything that one can, to become the most that one can be. Maslow believed that to understand this level of need, the person must not only achieve the previous needs, but master them, and only few persons can achieve this level.

To build the hierarchy of needs with the Unicoop data, the products have to be divided in classes, according to their PS value. To perform this division the authors use the ck-means alghoritm (an evolution of k-means algorithm) setting k = 5, following the Maslow's hierarchy of needs classification, in order to obtain the following classes of products: fundamental for survival, basic needs, complementary needs, accessory needs and luxury needs. The results of the ck-means clustering have been depicted in Figure 3.4 where, for each level of the hierarchy its main composition it's reported according to the product categories. The share values between the parenthesis tell, given the total amount of products purchases at that level of the hierarchy, how many of those belong to that particular category.

The authors report only categories representing at least 2% of the hierarchy level. They did not report the single product segment, as they are too specific and too many: for instance apples, pears, bananas, tomatoes, potatoes and so on have been aggregate in the product category Fruits & Vegetables.

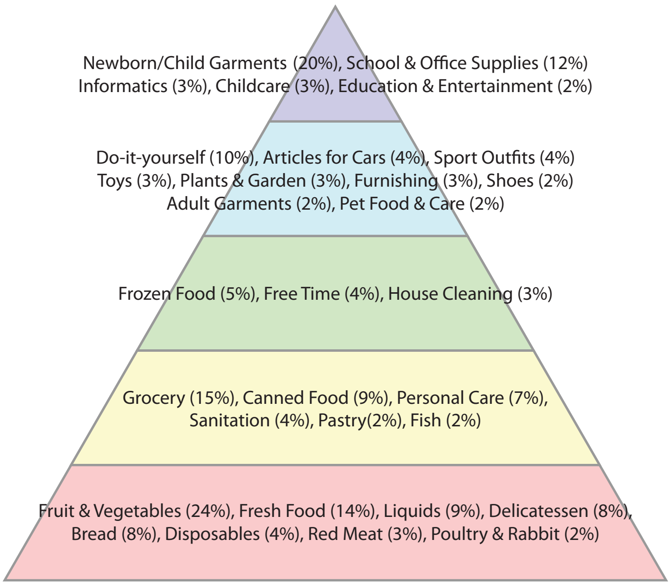
Newborn/Child Garments (20%), School & Office Supplies (12%)
Informatics (3%), Childcare (3%), Education & Entertainment (2%)
Do-it-yourself (10%), Articles for Cars (4%), Sport Outfits (4%)
Toys (3%), Plants & Garden (3%), Furnishing (3%), Shoes (2%)
Adult Garments (2%), Pet Food & Care (2%)
Frozen Food (5%), Free Time (4%), House Cleaning (3%)
Grocery (15%), Canned Food (9%), Personal Care (7%),
Sanitation (4%), Pastry(2%), Fish (2%)
Fruit & Vegetables (24%), Fresh Food (14%), Liquids (9%), Delicatessen (8%),
Bread (8%), Disposables (4%), Red Meat (3%), Poultry & Rabbit (2%)

Figure 3.4: Obtained Pyramid of Needs with the most basic products at the bottom and the most sophisticated at the top. For each product category we report its share among all purchase at that level of the hierarchy.

Of course, products in the same category may fall in different hierarchy levels: in Figure 3.4 they chose to put the category where it occupies the largest share of the level purchases. Figure 3.4 is telling some expected and some unexpected things. The basis of the pyramid is expected: most basic needs are food and personal hygiene. The top of the pyramid is instead telling us something surprising. Traditionally, reproduction is considered one of the most basic needs of any living thing. However, what we see is that in our modern society to have a baby ends up being one of the most sophisticated need, and the first one to be dropped, even before having a pet.

3.1.4 Marketing

Now that we talked about the hierchy of the needs, we will introduce a way to apply the Economic Compexity results to the marketing function.

Suppose the supermarket wants to promote a product p_i and it wants to limit its target to the smallest subset with the highest probability of actually buying the product advertised. The f_* function can be used in the following way: given the amount of products bought by customer c_j we use its index j to obtain the index $f_*(j) = i$ of the most sophisticated product p_i that c_j is buying, and therefore the entire set of products he/she is expected to buy, that is $assortment(c_j)$, defined as all the products that have an index $i' \leq i$. The same applies considering as input a product p_i, we obtain the index delimiting the set of customers buying it (for which $j' \leq f_*^{-1}(i)$).

One concern needs to be addressed before continuing: how well is the f_* function dividing the ones from the zeros w.r.t. what we expect? How much is a customer more likely to buy a product following the f_* function evaluated on our real world data (P_f) over any random product (P)?

The purchase matrix has ∼37 millions ones out of ∼1.5 billions cells, thus given a random product p_i and a random customer c_j the baseline probability $P(p_i, c_j)$ that customer c_j is buying product p_i in a significant amount (i.e. $\text{lift}(c_j, p_i) > 1$) is the ratio of these two numbers, or $P(p_i, c_j) = 2.44\%$. If we consider only the portion of the matrix at the left of the calculated isocline, i.e. the area of the matrix for which the f_* function tells us that the customers are very likely to buy exactly that products, we count 16,748,048 ones and 60,025,000 total cells. Thus, the probability $P_f(p_i, c_j)$ for a customer c_j to buy significant amounts of a product p_i for which $i \leq -\dfrac{\alpha j + \delta}{\gamma j + \beta}$ (i.e. $p_i \in assortment(c_j)$) is 27.9%. Using the f_* function, we can narrow of two orders of magnitude the set of combinations of products and customers to analyze and still capturing almost half of the significant purchases. In other words, customers are 11.43 times more likely to buy a product p_i if i is lower than, or equal to, the index limit predicted by the f_* function.This ratio is referred as $\frac{P_f(p_i,c_j)}{P(p_i,c_j)}$, i.e. the f_* function based probability of connecting customer c_j with product p_i over the baseline probability. They also calculated the same ratio, this time by counting at the right side of the isocline, where they expect to find many

p_i	p_{i-1}	$P(p_i)$	$P(p_i\|p_{i-1})$
Dishwasher Salt	Dishwasher Soap	8.39%	30.41%
Asparagus	Olive	8.00%	26.12%
Peppers	Chicory	7.31%	23.73%
Canned Soup	Preserved Anchovies	9.96%	32.23%
Wafers	Sugar Candies	11.30%	21.67%

Table 3.1: The probabilities of buying product p_i in general ($P(p_i)$) and given that a customer already buys product p_{i-1} ($P(p_i|p_{i-1})$).

zeros. The number of ones is 37 millions minus 16 millions, and it is divided by the number of cells, 1.5 billions minus 60 millions. The probability of obtaining a one is 1.39%, less than one twentieth of the left side of the isocline.

Now that the authors have addressed the main concern about the f_* function, they can safely assign to product p_i a corresponding customer index $j = -\dfrac{\beta i + \delta}{\gamma i + \alpha}$ that is its current "border": all indexes $j' \leq j$ represents customers who buy product p_i (i.e. $\forall j' \leq j, c_{j'} \in \mathit{customer\_base}(p_i)$), while the indexes $j'' > j$ are customers not buying p_i. By definition, the higher the value of j'', the more unlikely is the customer buying p_i. Thus, the set of customers the law is suggesting to target is the one immediately after index j. Since the f_* function is an interpolation, it is safe to define a threshold ϵ_1. Then, we define the set TC, the target customers set, as the set of all customers for which, given their index j', it holds: $j - \epsilon_1 \leq j' \leq j + \epsilon_1$ and $M_{cp}(c_j, p_i) \neq 1$ (the last condition is necessary to exclude from TC all customers who are already buying large quantities of product p_i, as it is useless to advertise p_i to them).

How can we evaluate how many elements of TC will be likely to start buying p_i? Iit holds that having a 1 in the product of index $i - 1$ makes the customer very likely to buy the next more sophisticated product p_i, i.e. to have purchased large amounts of the product immediately to the left in the matrix to p_i increase to probability of purchase this product. For

p_i	$\lvert TC^*\rvert$	$\lvert TC\rvert$	$\frac{\lvert TC_r\rvert}{\lvert TC\rvert}$
Tomino Cheese	58	137	7.51095
Raw Ham	78	144	5.81250
Apricot Jam	66	127	4.66142
Anchovies	83	144	4.06250

Table 3.2: The comparison between the size of the target customer sets identified by the f_* function against random target customer sets with the same number of customers likely to buy p_i.

instance, customers buying "Dishwasher Soap" have 30.41% probability of buying product "Dishwasher Salt" against a baseline probability of 8.39%, some instances of this are provided in Table 3.1. On average, the $\dfrac{P(p_i|p_{i-1})}{P(p_i)}$ ratio is 1.993 for the 500 most sold product, and no single product has a ratio lower than 1 (the lowest is 1.05 for Fresh Bread). Therefore, for each $tc \in TC$ element we check if $\exists x, M_{cp}(tc, p_x) = 1$, with $i - \epsilon_2 \leq x < i$, thus looking not only at the direct left neighbor of product p_i, but at his ϵ_2 left neighbors. If the condition holds, TC^* has been identified as the subset of TC composed by those customers who are likely to buy p_i.

The question now is: how large should be a TC_r set to obtain an equally large TC^*_r set if TC_r has been populated without knowledge about the f_* function, i.e. at random by picking customers who are not already buying product p_i? The authors address this question by looking at several different products. For each of them they identified the TC set using the f_* function and then they calculated 500 random TC_r sets. In Table 3.2 they report, for each product p_i, the following statistics: the number of customers likely to purchase p_i ($|TC^*|$ column), the total number of targeted customers ($|TC|$ column) and the average ratio between the targeted customers without and with using the f_* function ($\frac{|TC_r|}{|TC|}$), by fixing $\epsilon_1 = 100$ and $\epsilon_2 = 2$. As we can see, the knowledge provided by the f_* function reduces the number of customers to be targeted by a marketing campaign of four or more times, with the same return of investment (as our procedure fixes $|TC^*| = |TC_r^*|$). Table 3.2 reports only a few products, but they tested these 500 random sets for 800 different products and the average of the averages of the $\frac{|TC_r|}{|TC|}$ ratio is 3.55594, i.e. on average using the f_* function the marketing campaign can target three times less customers or less. For none of the 800 products the average of the ratio was less than 1.

3.1.5 Range Effect

In [12] the authors used the same method described above to obtain the products sophistication, but they used it for different pourposes. They want to show that the more sophisticated is the need to satisfy, the more cost the customers are willing to pay, in terms of distance to travel, to buy the product that can satisfy that need. They call this effect *range effect* of products.

It's possible to find products for wich customers traveled more than 5 kilometers on average, other products for wich average distance is less than 1 kilometer and many other products generated a variety of average distances. There are two trivial explanations on this fact: it is driven by price and/or by the frequency with which a product needs to be purchased. It is expected that customers will travel more to purchase products that are more expensive, for many possible reasons (they require higher quality, they may be not available around them, and so on). To check this hypotesis, first of all, given a price, they averaged the distance traveled by the customers buying the products with that exact price. Buy averaging, the ability of describing each single customer is lost, but there is the possibility to describe the behavior of the sytem in its entirely. The resulting is depicted in Figure 3.5: the price is on the x axis (in logarithmic scale), while the distance traveled is on the y axis. The price is recorded in Euros. Each dot is a purchase and we color it accordingly to how many purchases are represented by the same price and by the same distance.

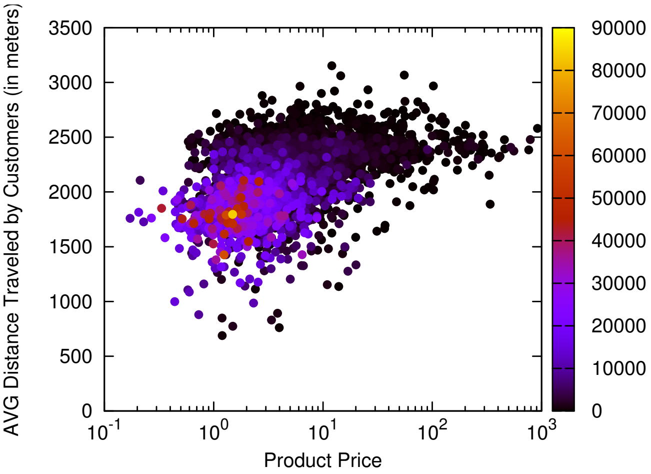

Figure 3.5: Average distance traveled to get a product with a given price.

From this figure a weak but positive correlation emerges, the price plays a role in driving customer decisions of traveling a given distance for a product. They calculate a log-normal regression2 using the function $f(x) = alogx + b$. In this regression, $R^2 = 17.25\%$, meaning that we can explain 17.25% of the variance in the distance traveled using the price.

The second hypothesis is that the frequency of purchase can have some correlation with the distance. We can suppose that a customer will travel a shorter distance to buy a product that he usually buy very frequently and this can happen for many reasons (to buy a product that is often needed, e.g. bread, the customer would prefer a shop very close to his home). To check if the frequency of purchase can explain the distance traveled by the customers, the authors repeated the same analysis using the number of purchases of a product instead of the price. From Figure 3.6 we can see a negative correlation: the more frequently a product needs to be bought, the smaller the distance a customer will travel.

The regression calculated with the function $f(x) = alogx + b$ has $R^2 =$

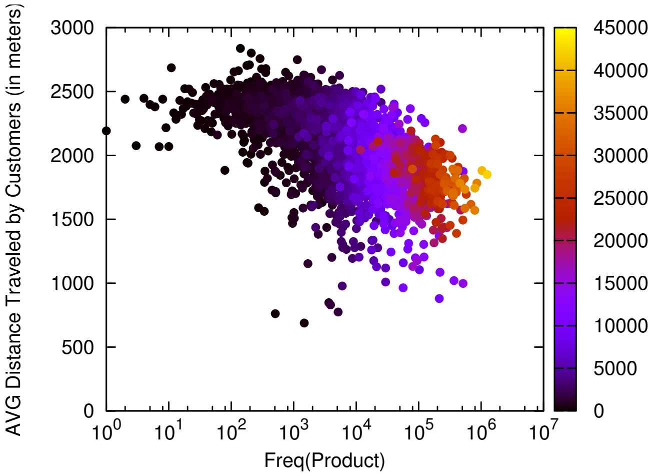

Figure 3.6: Average distance traveled to get a product with a given popularity.

32.38%

As conclusion the authors state that price and frequency of the need of a product play a smoll role in predicting the distance a customer will travel for purchasing a product, however, there is a large amount of variance that remains unexplained.

To better explain this variance, they use the products sophistication: their theory states that customers travel more to buy a product if the product can satisfy a more sophisticated need and/or they have sophisticated needs in general.

To validate this theory they depicted the product sophistication (x axis) against the average distance traveled by the customers to purchase the given product (y axis) in Figure 3.7.

The relationship is clear: the more a product is sophisticated, the more customers will travel to buy it. They calculated a linear regression, for wich $R^2 = 85.72\%$, explaining much better than previous tries the variance in the distances traveled by customers.

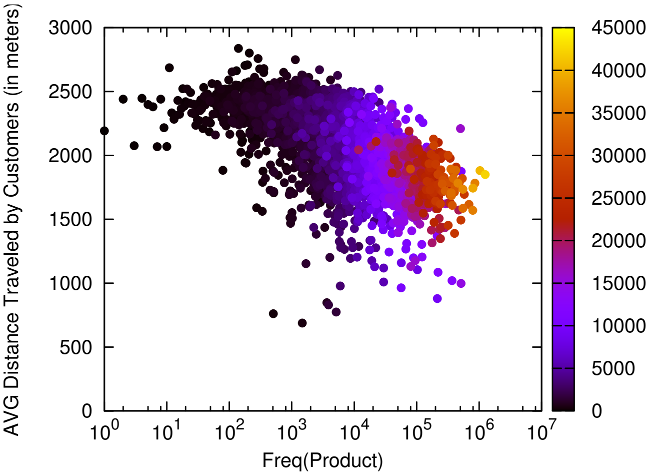

Figure 3.7: Average distance traveledby a customer with a given sophistication index.

In this section we explained how the authors use large sales data to extract information from the properties of these data. These properties are the uneven distribution of connections in the customer-product bipartite graph and the triangular structure of its adjacency matrix. They found that there are products that almost all customers buy (basic and sophisticated ones), while more sophisticated products are only bought by customers buying everything. These evidences lead to many applications: the hierarchy of needs based on retail market data, some marketing applications and the product range effect.

In the next section we will present the macroeconomic approach of Economic Complexity method.

3.2 Macroeconomic Approach

In [11] the authors assert that the productivity of a country residers in the diversity of its available nontradable capabilities and therefore, cross-country differences in income can be explained by differences in economic complexity, as measured by the diversity of capabilities present in a country and their interactions. In this work they present a method, called *Method of Reflections*, applied to trade data, to extract relevant informations about the availability of capabilities in a country. They interpret the variables produced by the *Method of Reflections* as indicators of economic complexity and show that the complexity of a countrys economy is correlated with income and that deviations from this relationship are predictive of future growth, suggesting that countries tend to approach the level of income associated with the capability set available in them. In the end, they show that the level of complexity of a countrys economy predicts the types of products that countries will be able to develop in the future, suggesting that the new products that a country develops depend substantially on the capabilities already available in that country.

In the *Method of Reflections* they rapresent the network composed by countries and products by the adjacency matrix M_{cp} where where $M_{cp} = 1$ if country c is a significant exporter of product p and 0 otherwise. They consider country c to be a significant exporter of product p if its *Revealed Comparative Advantage (RCA)* (the share of product p in the export basket of country c to the share of product p in world trade, **pratically the same concept of lift**) is greater than some threshold value. The method consists of iteratively calculating the average value of the previous-level properties of a nodes's neighbors and is defined as the set of observables:

$$k_{c,N} = \frac{1}{k_{c,0}} \sum_{p} M_{c,p} k_{p,N-1}$$

$$k_{p,N} = \frac{1}{k_{p,0}} \sum_{c} M_{c,p} k_{c,N-1}$$

for $N \geq 1$. With initial condizions given by the degree of countries and products:

$$k_{c,0} = \sum_p M_{c,p},$$

$$k_{p,0} = \sum_c M_{c,p}.$$

$k_{c,0}$ and $k_{p,0}$ represent, respectively, the observed levels of diversification of a country (the number of products exported by that country), and the ubiquity of a product (the number of countries exporting that product).

The authors provide empirical evidence that the method of reflections extracts information that is related to the capabilities available in a country by looking at a measurable subset of the capabilities required by products. They found a strong positive correlation between the average number of employment categories going into the export basket of countries and the family of measures of diversification provided by the method. This shows that more diversified countries indeed produce more complex products, in the sense that they require a wider combination of human capabilities, and that $\vec{k_c}$ is able to capture this information. They also show a correlation between $\vec{k_c}$ and income. Deviatons from the correlation between $\vec{k_c}$ and income are good predictors of future growth, indicating that countries tend to approach the levels of income that correspond to their measured complexity.

This method was lately revised by Caldarelli [14] cause it was proven that, under some circumstances, it may not converge. The main change proposed by Caldarelli is the initial condition, that is the inverse of the degree of countries and products:

$$k_{c,0} = \frac{1}{\sum_p M_{c,p}},$$

$$k_{p,0} = \frac{1}{\sum_c M_{c,p}}.$$

In this chapter we discussed about the applications of economic complexity in a microeconomic and a macroeconomic environment In next chapter we will discuss about the dataset used to follow up the microeconomic approach to give a reasonable measurement of human well being.

Chapter 4

Data

Now that we exposed the existing methods to calculate the Economic Complexity, we will talk about our dataset, providing a description of it and the problems inherent data selection and preparation.

4.1 Dataset Description

The dataset was provided by Unicoop Tirreno, one of the largest Italian retail distribution company, and contains all purchases of customers between 2007-01-01 and 2012-12-31.

In the data warehouse, the fact that we studied is the single scan of the product, that provides informations about the customers, the products they buy, the date and the place of the purchase. The conceptual model of the data warehouse is depicted in figure 4.1.

An important dimension of the data warehouse is Marketing, representing the classification of products: it is ogranized as a tree and it represent a hierarchy built on the product typologies, designed by marketing experts of the company. The top level of this hierarchy is called “Area” and is divided in three categories: “Food”, “No Food”, “Other”. The bottom level is called “Segment” and it contains $7,679$ different values (Table 4.1).

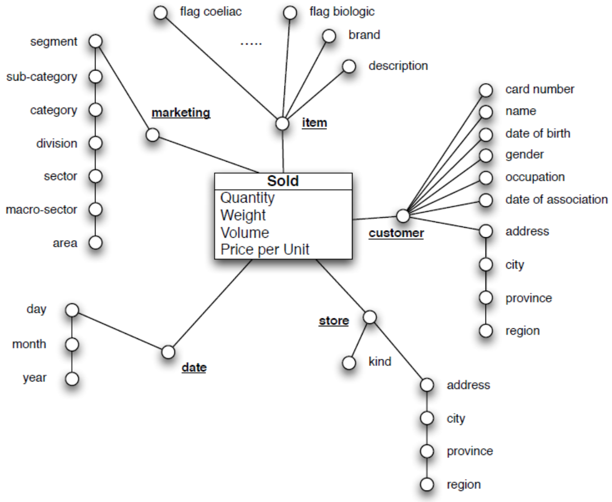

Figure 4.1: The data model of Data Warehouse.

Area	**Macro-Sector**	**Sector**	**Division**
3	4	13	73
Category	**Sub-Category**	**Segment**	**item**
529	2197	7679	514801

Table 4.1: Number of distinct elements for each level of marketing hierarchy.

The active and recognizable customers are $1,260,458$. A customer is active if he/she has purchased something during the data time window, while he/she is recognizable if the purchase has been made using a membership card. The 129 stores of the company cover the whole west coast of Italy, selling $514,801$ different items (Figure 4.2).

Since we want to build an adjacency matrix that represent the bipartite graph customers-products, we have to take the data from the sold table of the Data Warehouse. The resulting matrix $customersXproducts$ (M_{cp}), according the dimensions described above, would have $1,260,458$ rows and $514,801$ columns, but this kind of matrix would present computational and conceptual problems, so we need to choose some criteria to reduce dimen-

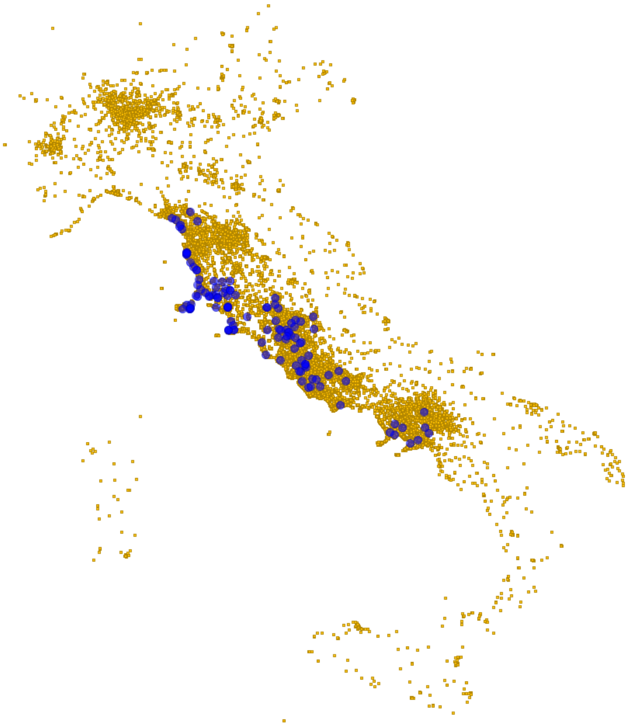

Figure 4.2: Distribution of customers (yellow) and shops (blue).

sions and keep only meaningful data.

4.2 Data Selection

The M_{cp} matrix needs to be reduced in some way, so we have to perform some filter on customers and products. In a first step of analysis, we decided to not apply any filter on customers except than the customer with $ID = 0$. This ID represents all the customers without a membership card, so, in practice, it includes multiple customers treated as one and we need to exclude them from our calculations. An strong filter, instead, was made on products. This selection was made to deal with two problems: first of all the huge dimension of the resulting matrix, and second, the excessive level of detail, that drives us at loosing some important informations. Indeed, considering the following products

- Pepsi Cola 1.5L
- Coca Cola 1.5L
- Coca Cola 0.5L
- 6X Coca Cola 1.5L
- 6X Coca Cola Christmas Edition

we have all different items, but, for our purpose, we should consider them as a only one item. Generally speaking, we would like to aggregate products that are similar, without considering differences in packages, size and brand.

In order to solve this problem, we chose to use the marketing hierarchy described in table 4.1. We decided to sobstitute the item with the value of the marketing Segment, in this way che cardinality of the dimension of the products was reduced by 98% (from $514,801$ to $7,128$), aggregating at the same time products that are equivalent (for example, at this level, we identify with "Sugar Free Orange Juice" both the liter and half liter bottle items cause we are not interested in distinguishing the different packaging of the same product). Once we decided the granularity to be used, we had to exclude from the final matrix that segments that are meaningless for our analysis, for example shoppers, discount vouchers, errors, segments never sold, etc. Finally, we have our products (segments) pool to calculate the Sophistication. Obviously, this selection caused the discharge of the customers that bought exclusively products classified into the removed segments. The final step was to choose the time granularity: the doubt was between using a monthly aggregation or a quarterly aggregation. We chose to use a quarterly aggregation mainly because we wanted to compare our results with GDP, and GDP assume a better relevance in a quarterly aggregation. For each quarter, we have $\sim 550k$ of active customers. Aggregating in quarters, we noticed that the first two years of our dataset (2007-2008), have a great

1st quarter 2007	**2nd quarter 2007**	**3rd quarter 2007**	**4th quarter 2007**
10776	11952	17507	15471

1st quarter 2008	**2nd quarter 2008**	**3rd quarter 2008**	**4th quarter 2008**
17641	38591	109851	456498

1st quarter 2009	**2nd quarter 2009**	**3rd quarter 2009**	**4th quarter 2009**
499738	547947	610596	521547

1st quarter 2010	**2nd quarter 2010**	**3rd quarter 2010**	**4th quarter 2010**
502768	546301	613094	523071

1st quarter 2011	**2nd quarter 2011**	**3rd quarter 2011**	**4th quarter 2011**
502221	547699	619335	524291

1st quarter 2012	**2nd quarter 2012**	**3rd quarter 2012**	**4th quarter 2012**
501850	541332	627035	528568

Table 4.2: Number of distinct customers for each quarter

difference, in terms of data cardinality, with the next years and, probably, this is due to the fact that the data collection by Unicoop was in a primordial state (Table 4.2).Due to this difference, we decided to exclude 2007 and 2008 from the analysis.

In a second step of analysis we had to calculate the customers sophistication. In this phase we had to answer to a question: how much infrequent customers can affect our analysis in comparing different quarters?

In order to avoid this problem, we selected a restrict pool of customers, precisely the customers who went to the shop at least three times in every quarter. So, while for product sophistication was meaningful to include all customers, for customer sophistication we applicated this filter and reduced customers by 86% (from $1,260,458$ to $172,884$).

In this chapter we described the used dataset and discussed about its main dimensions. We talked about the problem of filtering and adjusting data in order to obtain meaningful matrices and the issue of aggregate them in quarters. In the next chapter we will report the application of Economic Complexity method to this dataset and we will show our results.

Chapter 5

Experiments

The previous chapter introduced the dataset that we used for our analysis, we discussed about the main dimensions and the selection of the data. In this chapter we will describe the exeriments we did, starting from the first and simple ones, up to the application of the Economic Complexity method. At the end we will show all the results of the Economic Compexity method, in a graphical and statistical way.

5.1 Simple Aggregations Analysis

The objective of our work is to find out a way to approximate human wellness, providing realistic and almost “real time” results, following the idea at the basis of the concept of nowcasting introduced previously.

In a first step of analysis we want to see if some simple aggregation of data could give us some good information about the trend of GDP. The easiest correlation that we can try to check is the one related to the amount spent by customers in a quarter and the GDP value in the same quarter. GDP, as we said, can be calculated with the sum of all expenditures incurred by individuals in one year and we tried to repropose something similar to this method summing all expenditures of Unicoop Tirreno customers in a quarter. The first problem of this approach is that, in some quarters, we can have more or less customers than in the other ones (for example, during summer, some customers can be away from their region for holydays) and this make difficult to compare different quarters. The solution was to average the total amount of expenditures in a quarter by the number of customers who purchased something in that quarter. To make comparable this value with GDP value, we normalized both in a scale between 0 and 1.

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

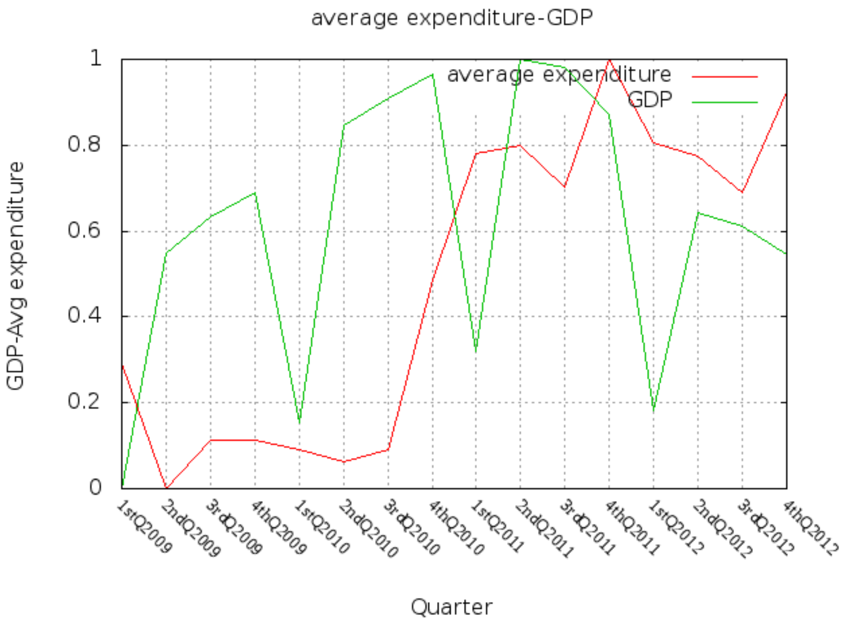

Figure 5.1: GDP and average expenditure trends. Values are normalized between 0 and 1

In Figure 5.1 we can see the two trends compared. The evidence is that there isn't any correlation between the two values: we cannot approximate GDP with the average of expenditures of the customers.

The second analysis is to compare GDP trend with the number of items purchased in each quarter.

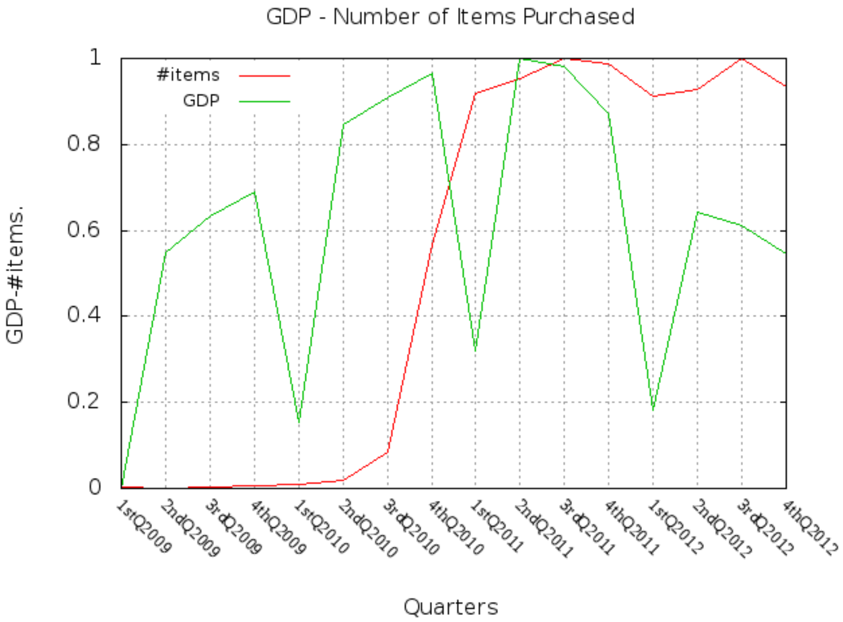

Figure 5.2: GDP and number of items purchased. Values are normalized between 0 and 1

Figure 5.2 doesn't show any correlation between the two values.

5.2 Economic Complexity in Quarters

After we found that a simple sum of expenses cannot be meaningful, we calculeted the Economic Complexity Index for products and customers for each quarter.

The first step was to build the bipartite graph starting from the Sold table of our Data Warehouse. For this pourpose, for each quarter, we created a table with three rows: *CUSTOMER_ID, PRODUCT_ID,#ITEMS*. In this table we have a customer c_j, a product (segment) p_i and the amount of product p_i purchased by customer c_j. The query to build this table was:

	$product_i$	....	$product_m$
$customer_1$	$lift_{1,1}$	....	$lift_{1,m}$
	....	....	....
$custmer_n$	$lift_{n,1}$	....	$lift_{n,m}$

Table 5.1: Lift matrix.

```
SELECT customer_id, marketing_id, count(*)
FROM sold s
WHERE s.data_id between “start quarter” and “end quarter”
GROUP BYcustomer_id, marketing_id
```

Starting from the resulting table, we built the matrix containing, for each couple (c_j, p_i), the $lift(c_j, p_i)$ (Table 5.1) and from this matrix we are able to calculate the product sophistication index, with the method described in section 3.1.

What we get are 16 vectors one for each quarter in the period [2009/1/01 - 2112/12/31] containing the sophistication index of each product sold in that quarter. For the calculation of customer sophistication we couldn't use the algorithm because, in order to calcolate the sophistication index for a partition $\mathcal{P}$*(Product or Customer)*, the method requires to build a matrix $\mathcal{P}X\mathcal{P}$. While for the products the dimensions are acceptable ($\sim 6,000X6,000$), for the customers partition, the matrix could be not-manageable ($\sim 550,000X550,000$). Due to this, another way to calculate the customers sophistication is needed. The first option we thought was to calculate, for each customer, in each quarter, the average of sophistication of the products that he purchased in this quarter, but, to do this, we had to see the distribution of this products to see if the average was a meaningful function of aggregation.

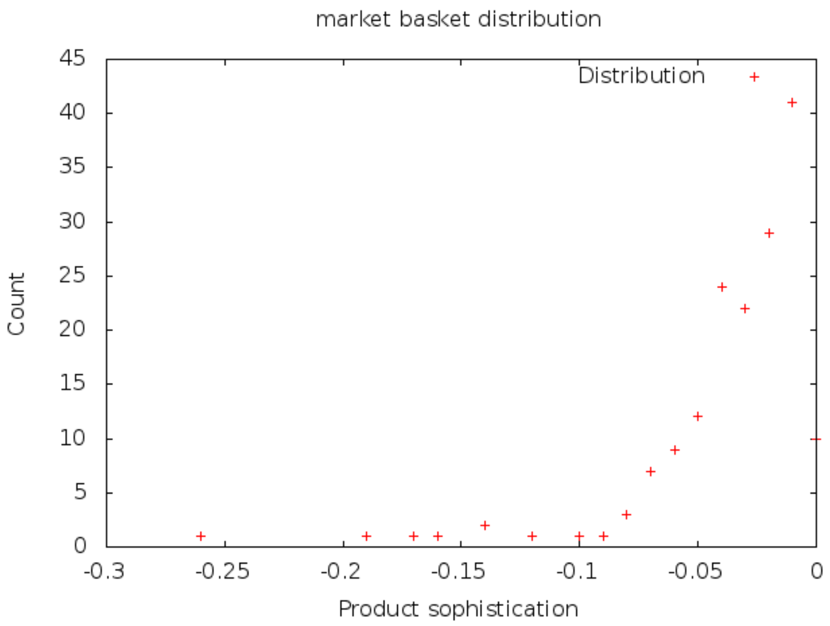

Figure 5.3: The distribution of the sophistications of the products bought by a cutomer in a quarter

As we can see in Figure 5.3, this distribution seems to be a power low, so we decided to not use the average to calculate the customers sophistication, but we used the median as aggregation function: the median of a set of products P = (p_1,...,p_i) is defined as follow:

$$Median(P) = \begin{cases} = \frac{p_{\frac{i}{2}} + p_{\frac{i}{2}+1}}{2} & \text{if i\%2 = 0} \\ = p_{\frac{i+1}{2}} & \text{otherwise} \end{cases}$$

So we calculated, for each quarter, the sophistication of each client active in that quarter.

5.3 Studying Products Sophistication

At this point we calcuated the products sophistication for each quarter in the time period considerated, so we need to describe the studies over the obtained results. First of all, we have to find an aggregation function to summarize the sophistication of all products in each quarter. Again, we plotted the probability distribution of products sophistication for each quarter, in order to choose the function we can use.

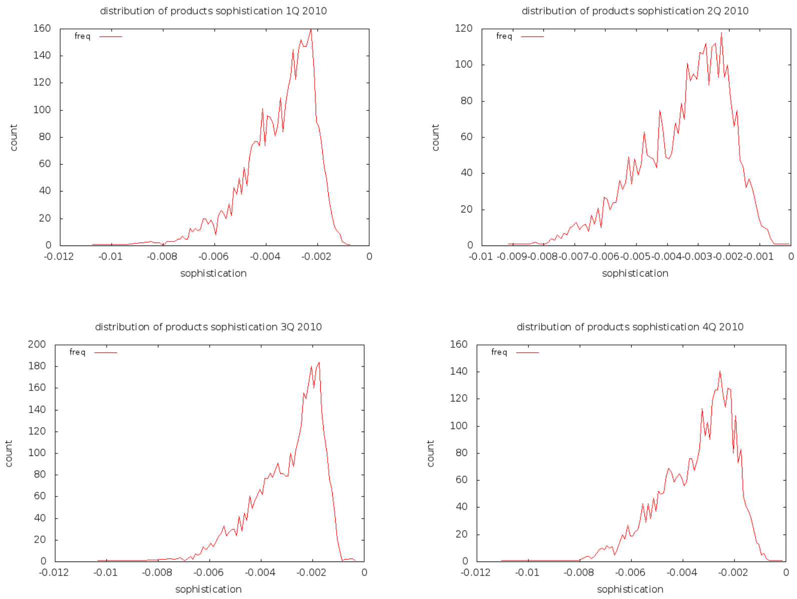

Figure 5.4: Distribution of the products sophistication in 2010 quarters

In Figure 5.4 we report the whole 2010 as sample of distribution. This example is about 2010, but we can find the same distribution in all the other years. We can see that the distribution has a normal form, so we can aggregate the products sophistication using the average for each quarter. Now we need to compare these values with GDP values. We normalized both GDP and sophistication values to have the same order of magnitude and make them comparable. We used the same formula used above:

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

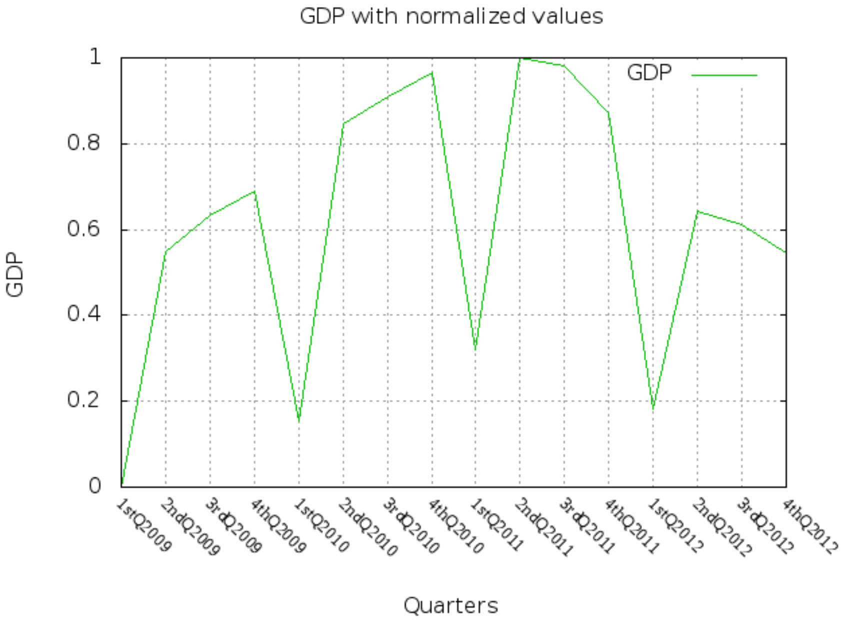

Figure 5.5: GDP trend: the values are normalized between 0 and 1

In Figure 5.5 we show the trend of the GDP. We can notice a seasonal component, going down on the first quarter of each year. GDP is available in both versions, seasonally adjusted or not adjusted, and we had to evaluate wich one to use. GDP is seasonally adjusted by a method called *X-12-Arima* that is developed by the U.S. Census Bureau[15]. To make the two series comparable, we should adjust our data with the same algorithm, but this procedure has two problems: 1) the method is based on the use of a moving average that needs a larger time window than the one we used, and 2) he method can consider a large set of parameters defined by user, such us holiday effects, but we are not sure about the effects of these parameters on sophistication, so, in case we include the example of holiday effects amd other similar parameters, we can lose some important information on sophistication trend. Considering this problems, we decided to use the not seasonally adjusted data for GDP.

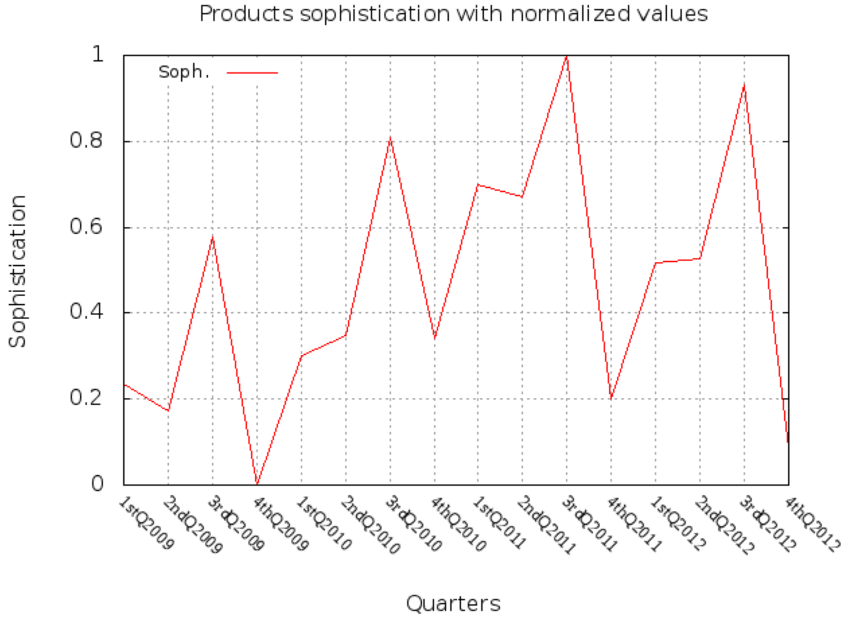

Figure 5.6: Products sopthistication trend: the values are normalized between 0 and 1

In Figure 5.6 we depicted the products sophistication trend. As we can see, it show the same seasonal component of GDP, but going down on 4th quarter of each year.

In Figure 5.7 we plot the two trends together, to show the correlation between the two values. Obviously, we had to give a statistical validation to the visual evidence. To do this, we used the *Pearson product-moment correlation coefficient*. This coefficient is a measure of the linear correlation between two variables X and Y, and assumes values between +1 and -1 inclusive, where 1 is total positive correlation, 0 is no correlation, and -1

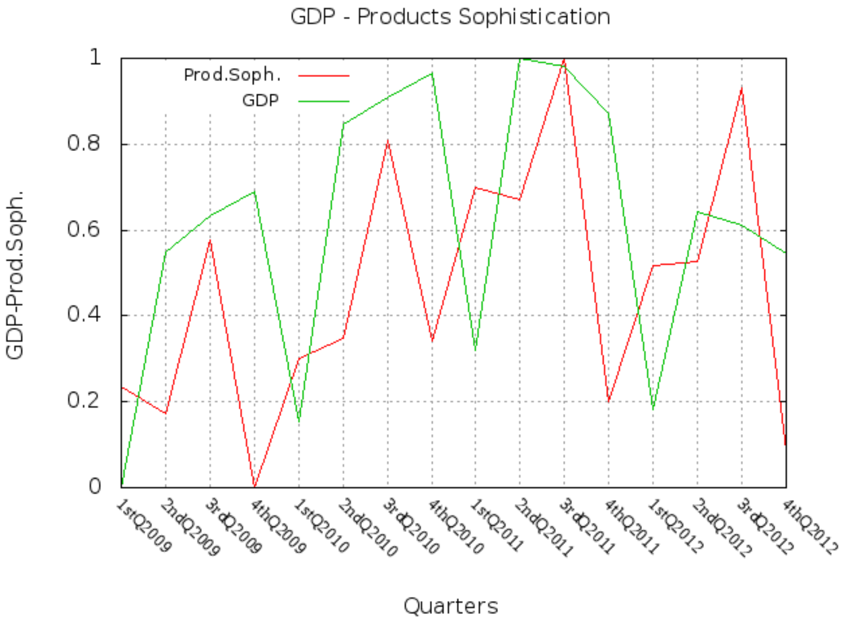

Figure 5.7: GDP and products sophistication trends: the values are normalized between 0 and 1

is a negative correletion. It is widely used in the sciences as a measure of the degree of linear dependence between two variables. It was developed by Karl Pearson from a related idea introduced by Francis Galton in the 1880s **??**. Pearson's correlation coefficient between two variables is defined as the covariance of the two variables divided by the product of their standard deviations, in formula:

$$P_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

Considering that product sophistication trend seems to be predictive compared to GDP, we calculated the Pearson coefficient between sophistication and GDP values translated of one quarter later. We got a *P coefficient* = 0.654, that is a very good value. Unluckly, the data we used to find this correlation are still incomplete in 2010.

5.4 Studying Customers Sophistication

while the products sophistication gave us good results, we want, for completeness, study also the customers sophistication. As described above, for each customer, it was calculated his/her sophistication index by averaging the sophistications of his/her whole basket with $lift > 1$ in each quarter. To choose the aggregation function for this sophistication, we depicted the distribution of the customers sophistication for each quarter, and there is an example in Figure 5.8.This example is about 2010, but we can find the same distribution in all the other years.

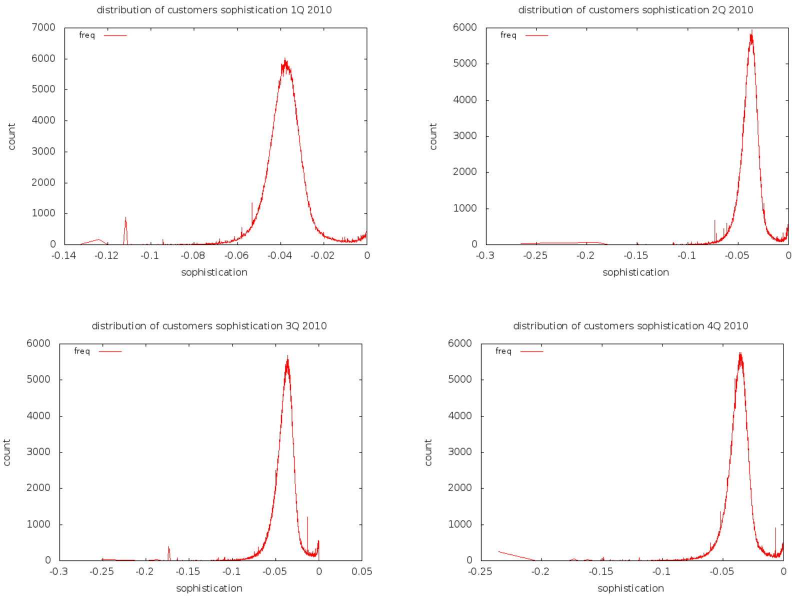

Figure 5.8: Distribution of the customers sophistication in 2010 quarters

We can see that the distribution has a normal form, so we can aggregate the customers sophistication using the average for each quarter.

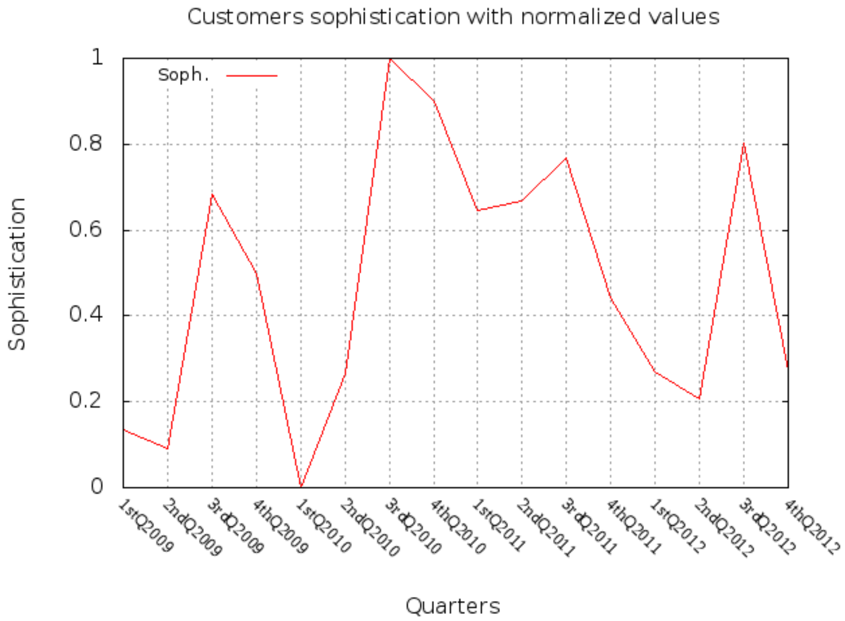

Figure 5.9: Customers sopthistication trend: the values are normalized between 0 and 1

In Figure 5.9 we depicted the customers sophistication trend. We compare this trend with the GDP trend in Figure 5.10

We calculated the Pearson correlation translating the GDP values one quarter later and the P correlation is 0,0785, so there isn't any appearent correlation between the two trends. This low value for the correlation GDP - customers sophistication may be caused by the two level of aggregating data in calculating the trend and for the filter on the customers that made us discharge of data.

We also tried to find some correlation between the trend of the variance of customers sophistication in each quarter and the GDP. This correlation may be existing because, growing up the GDP, may grow up inequality between customers.

Calculating the Pearson correlation between variance and the GDP, we found another good result, if confirmed with future analysis. Figure 5.11

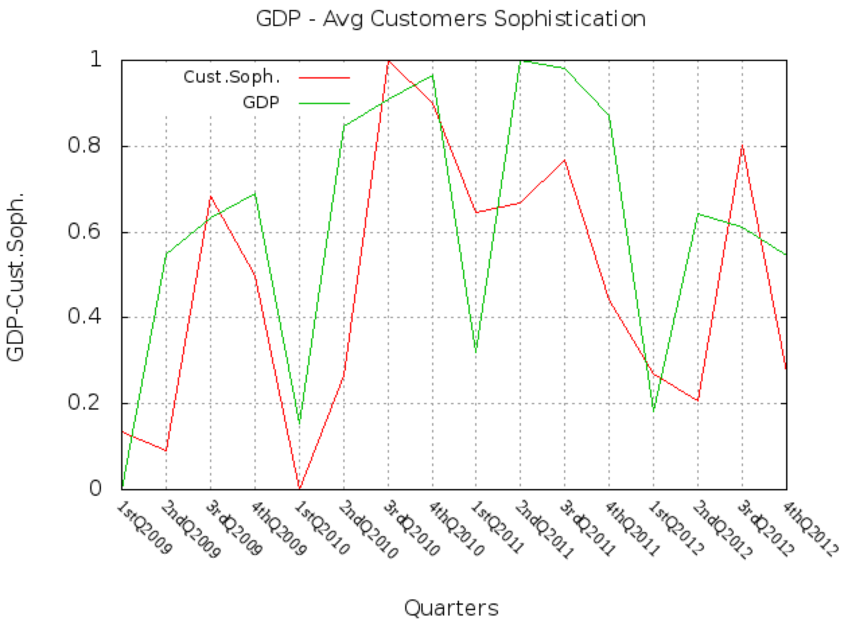

Figure 5.10: GDP and customers sophistication trends: the values are normalized between 0 and 1

shows a negative correlation, infact the Pearson correlation between the two trends is -0.4998. We must remember that a negative correlation is still a correlation. Growing GDP, inequalities among persons goes down. This result, if confirmed with future analysis, can be very important to predict the social inequalities of a Country. It means that, the less the sophistication level of the overall population is, the more the "rich get richer" effect becomes strong, e.g. the differences between people with high sophisticated and low sophisticated needs grows up. Obviously, this is not a general result, but it can depend on various factors, such as the historical period, the geographic area, the economic policies, and so on.

In this chapter we described our experiments over the Unicoop data. We first made some simple aggregation trying to find a correlation with the GDP.

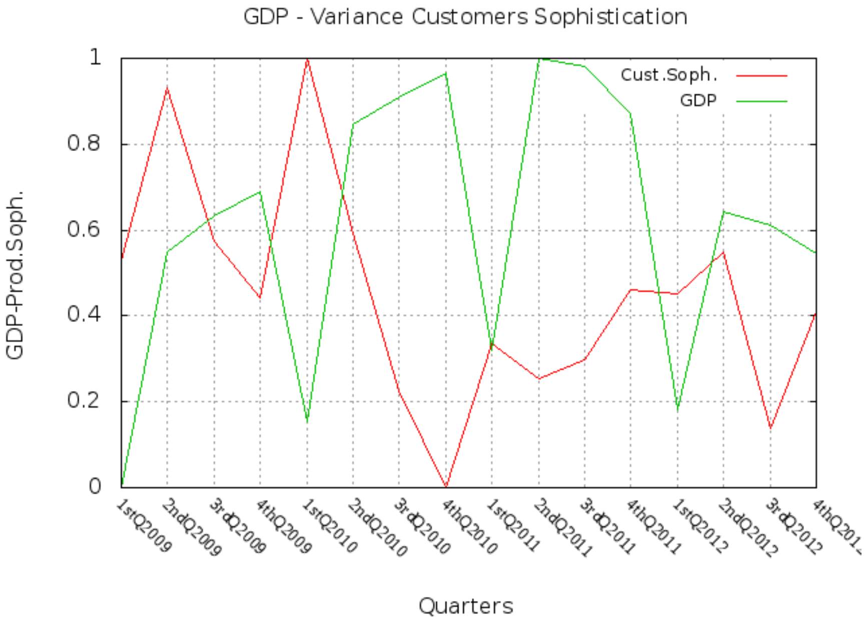

Figure 5.11: GDP and variance of customers sophistication trends: the values are normalized between 0 and 1

Then we tried to compare the average of the amount that the customers spent each quarter with the GDP, but theres'nt any apparent correlation, so we switched to the sophistication.

We calculated the products and customers sophistication using the sales data, represented by a bipartite graph customer-product and its adjacency matrix M_{cp}.

We compared the sophistication index of the products with the GDP trend and we found out a good correlation. This means that the sophistication trend forecasts the GDP trend. This correlation is good, but surely could be better with the use of a larger dataset. The dataset that we used contains the data of only four italian regions, while the GDP data we used is calculated with data of the whole Italy. Moreover, we didn't consider some kind of goods, such as luxury goods, and we consider only a subset of customers. On the other hand, our index does not require all the resources needed to calculate GDP. Our data are available almost in real-time, day by day, and we don't need to overhaul our results one, two or even more years later. As last, remembering the criticism over GDP, we have to say that our index is not intended to be exactly equal to the GDP, but is a way to extimate the human wellness. Human wellness can be approximated by GDP, but can never be equal to it, and so it is for our resulting index.

For completeness, we compared even the sophistication index of customers with the GDP trend. Using the customers sophistication we loose the predictive power that we had with the products sofistication, infact, comparing the sophistication and GDP trends, we found out no correlations. This is due to the addictional level of aggregation that we used to calculate the customers sophistication index and the discharge of data in filtering the customers. We believe that, having the possibility to calculate that index using the direct method described in this thesis, the correlation could be much better.

The final step of our analysis was to compare GDP trend with the variance of the customers sophistication. As we saw, the correlation was negative, that means that social inequalities (measured in terms of customers sophistication) grew up whenever the GDP went down. Is important to notice that this result holds for the particular context of our data (the area, the period, the socio-economic background, etc.), while it can change situation by situation.

The correlation was negative: growing GDP, inequalities among persons goes down. This result, if confirmed with future analysis, can be very important to predict the social inequalities of a Country. It means that, the less the sophistication level of the overall population is, the more the "rich get richer" effect becomes strong, e.g. the differences between people with high sophisticated and low sophisticated needs grows up. Obviously, this is not a general result, but it can depend on various factors, such as the historical period, the geographic area, the economic policies, and so on.

Chapter 6

Conclusions and Future Works

The importance of measuring human well-being is a central issue for politicians and statisticians. We explored the current methods to measure it and the limitations of these methods. One of the main criticism is the arbitrary choice of the variables needed in the calculation. We tried to overcome this problem with a method that is independent from this choice: to measure wellness of a population, we see what the people buy. We do not limit the analysis to the sum of the amounts spent by the customers, but we analyzed how they spent the money.

Another problem of the current methods is the wide use of resources to retrieve the data and, consequently, the strong delay with which this data is available for the analysis. The framework we used to measure the human well-being is able to avoid this problem cause it uses the data provided by a retail market company, that is available in real time, and, in this way, is possible to give an approximation of the state of health of the population in large advance with respect to the availability of results of current methods (in particular, we used GDP as a reference point). The concept of predicting the present, i.e. predicting results of some measurements before known methods can publish them, is known as *nowcasting.* The framework studies the properties of the bipartite graph customer-product to retrieve informations from the customers behavior. The graph is represented by its adjacency matrix that presents a triangular shape. This shape is the main concept at the basis of sophistication, it tells us that only few customers can satisfy high-level needs, and who satisfies them, has already satisfied all other needs. The framework assigns to each product a sophistication index and we can assign it even to each customer by averaging the index of the products that they buy: the more a product is sophisticated, the more is sophisticated the need that it satisfies and, the more a customer is sophisticated, the more sophisticated are his needs.

The sophistication indexes of both products and customers were calculated in quarterly aggregation, to compare their trends with GDP in a temporal evolution. We used the sophistication index cause the trend of GDP and the trend of other simple aggregations didn't give any good result. Compairing the trend of products sophistication with the trend of GDP, we found an high correlation and we measured this correlation by calculating the *Pearson correlation coefficient* between the two series. We also compared the sophistication of customers with the GDP. We found that there isn't a good correlation between the two trends. A possible explanation for this is that the customers sophistication derives from two aggregation levels, and is not the result of a direct calculation. At the end we found an anti-correlation between GDP and the variance of the a customers sophistication. This, at a first look, seems to underline that during periods with relative low GDP values, the inequality among customers in terms of sophistication grows up; and viceversa, with relative high GDP values. This should be object of further analysis, with an accurate statistical study to confirm our hypothesis.

The work of this thesis can be enriched int two different ways: first we can compare the trend of the sophistication with the trend of other newer indexes, that are able to show the wellness better than GDP. In addiction, can be performed not only a temporal analysis of the evolution of the sophistication index, but even a geographical analysis. The idea is to analyze the geographical distribution of the products and/or customers sophistication in a period t_0 and to study the evolution of this distribution in the following periods $t_1, ..., t_n$ to find some interesting patterns between geographical areas.

Bibliography

[1] Bureau of Economic Analysis (BEA), U.S. Department of Commerce, *Measuring the Economy: A Primer on GDP and the Nation al Income and Product Accounts* , september 2007

[2] Lawn, Philip A., *A theoretical foundation to support the Index of Sustainable Economic Welfare (ISEW), Genuine Progress Indicator (GPI), and other related indexes*, 2003

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

[4] Neumayer E., *On the methodology of ISEW, GPI and related measures: some constructive suggestions and some doubt on the threshold hypothesis*, 2000

[5] Lawn, Philip A., *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 (SNB). Environment, Development and Sustainability*, 2005

[6] Stephen M. Posner , Robert Costanza, *A summary of ISEW and GPI studies at multiple scales and new estimates for Baltimore City, Baltimore County, and the State of Maryland*, 2011

[7] Hyunyoung Choi, Hal Varian, *Predicting the Present with Google Trends*, 2011

[8] Domenico Giannone, Lucrezia Reichlinb, David Small, *Nowcasting: The real-time informational content of macroeconomic data*, 2008

[9] Kalman, R.E., *A New Approach to Linear Filtering and Prediction Problems*, 1960

[10] Diego Pennacchioli, Michele Coscia, Fosca Giannotti, Dino Pedreschi, *Discovering the General Patternof Shopping Behavior*

[11] Cesar A. Hidalgo, Ricardo Hausmann, *The building blocks of economic complexity*, 2009

[12] Diego Pennacchioli, Michele Coscia, Salvatore Rinzivillo, Dino Pedreschi, Fosca Giannotti, *Explaining the Product Range Effect in Purchase Data*

[13] Maslow, A.H., *A theory of human motivation*, 1943

[14] Guido Caldarelli, Matthieu Cristelli, Andrea Gabrielli, Luciano Pietronero, Antonio Scala, and Andrea Tacchella, *A network analysis of countries' export flows: firm grounds for the building blocks of the economy*, 2012

[15] `http://www.census.gov/srd/www/x12a/`

[16] Cave B.M. Lee A. Pearson K. Soper H.E., Young A.W. *On the distribution of the correlation coefficient in small samples*, 1917.