Getting familiar with microsimulation and Statistical Matching

I.1. Introduction

Let me introduce the microsimulation system: one of the head profile of our company, called Új Calculus Bt. Microsimulation is a method, which is able to modell the processes faithfully in theory. The micro word means the level of the method: all of the commands must be executed at the level of individuals of the population.

In these days we have to emphasize on predecision makings in order to stay competitive and get new customers. It is important to know the expected reactions of the customers before introducing a new product or change the conditions. Before the decision, a lot of parameters must be examined. The biggest problem is, that the institutes do not know their own customers enough. They have no up to date information about them. This information failure must be toned down with the usage of statistical tools, otherwise we can not forecast their reactions. Customers must be invested with real and actual consumption, demographic and economic properties.

We have a frame system, which attain the microsimulation. It is a software and we work with it.

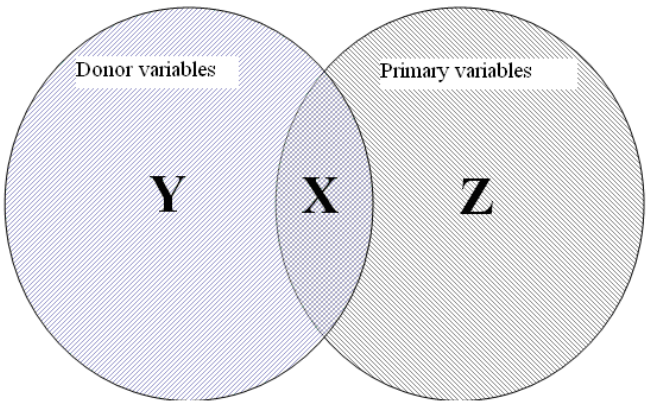
I.2. Microsimulation

The computational system transforms a data file by right of different hypothesises on micro level. With this method the properties(demographic or social, economic properties) of the population's individuals will be changed, so the population will also change. The main point of microsimulation is to take on the well known data file in function of time by the help of the computer. To take the data on, we need calculus of probability tools, regulations and algorithms, based on empirical reality. With the statistical analysis of the results, you can examine the effects of the assumed hypothesises and with these you can come to strategic decisions. One of the most important property of microsimulations, that you can get estimated data at low cost. Unless this simulation, you could get this data only with a new expensive field-day. Simulation can be extend for the changes of several years. The base of this possibility is the law, according to some group of the data file have the same properties.

I.3. Statistical Matcing

By the modelling of the effects of economic and social changes the most critical condition is to have corect data files to the simulation. To reach this goal, you can use the Statistical Matching method. In this process we join two data systems from different field-day, when there is no unique keys to join the two tables. So we have a primary data file, in which there are missing data and another one: the secondary(donor) data file, which we substitute the previous missing data from. The heart of the method is to look for couple from the data file on affinity not on keys to substitute the missing data in the other data file.

The Statistical Matching is a part of the data extension, where one case of one of the database is joined an other case with the similar property of the other database. The data extension has further meaning, than Statistical Matching. Every research and experiment, in which we create one database from two independent ones, are called data extension. Drawing it sipler: from two databases, which contain different cases, we ceate a new database thru the common variables. The place, where the variables are transported from is the donor database and the place, where the variables are fusioned from the donor database called primary database. The variables of the donor database, which will be fusioned, called y. The fusioned variables of the primary database, called z. The common variables are x.



The keys to the successful extension are:

- Two randomly choosen databases are not joined each other with a big likelihood. First we have to examine whether the two researches used the same population. If the two populations, which the elements of the sample was choosen from are not the same the extension could not be started.

- The common variables must be queried by the same way in both of the researches. If it is possible, the given answers should be almost the same. For example the highest school qualification is not equal to the highest school, you attended to, so these variables would not be good to be extension variables.
- If the common variables in both of the databases are the same, there would be the failure: the questioners did not behoove their works at the same level. To get correct linking, the two databases must be in the best quality.
- Listening to the time is important too. Both of the researches have to cover the same time or we have to assume, that the queried variables did not change between the two date of the data admissions. For example the religion of the pepole does not change for half a year, but their preferences to the political parties can be changed. It is worth to examine the distribution of the common variables. They have to match.
- To get a linking with good quality it is very important, that the x variables correlate with y and z variables strong.
- The last assumpion is the Conditional Independent Assumption(CIA). The extension is valid if and only if the CIA comes true. It means, that the y and z variables are independent from each other. There is not any coherence among y and z variables, which can not be explained by the variables x. If x variables are known, then y do not contain any further information about z and vice versa.

Statistical Matching originates in the problem of marriage: there is always marriage and the repeated selection is allowable.

II. Real applications of microsimulation

In the following part of this introduction some applications will be introduced. The most important usage of the microsimulation system is the demographic, social and economic effect of an action. With the help of this method, you can get a forecast about the effect of the action, before the decision of the government is codified. A lot of version of the edicts can be modelled. For example you can come to know the amount of the tax income and social effects by different tax steaks, which help the government in making decision.

II.1. Estimation research data from microcensus 2004

We took part in the microcensus in 2004. We made data correction on the data file about the income. (Some people did not answer the questions about the salary, so we had to substitute the missing data.) The Hungarian Central Statistical Office(HCSO) made the Home Expenditure Survey(HES). We used Statistical Matching method for these two data files and we got the research data, which contains typical households without any concrete data(for example name) because of the data protection law. You can not go back to the same households to ask them about the changes in their income and careers, so ontaking is necessary. The research data is unusable, if it is not ontook by demographic properties. The ontaking of income and career is ready until 2010, we are looking forward to the census for 2011 to compare our results with it. Now we make comparative analysis of different tax systems in different countries in Europe.

You can use microsimulation to take the data of the previous years on. There is not any census every year, because it is very expensive, but the HCSO has to behoove analysis every year. It is not possible to use the same data file for several years. The solution for this problem can be the ontaking of the demographic data. The ontaking of the demographic data means if we have a data file for 2009 and the changes in 2009 is available(for example how many people died, or how many children was borned) and we take these changes for this data file on, we get the data file for 2010. To get this result we have to count the death- and congenital probabilities out and use microsimulation with these probabilities, namely each of the people we have to decide whether she bears and if yes, what the properties are of the children(girl or boy, twins or no twins) or he/she is the choosen person, who will die.

We use microsimulation to modify a data file in micro level, thus modify at the level of person and household. If we would like to do anything for every people and households(for example increase the salary of each people), it is achievable without any problem, but if we would like to do something with only the part of the people, we can do it with the help of the probabilities. For example we can not increase the salary of 100 people, because microsimulation is not for this purpose. We ususally would not like to increase the salay of a given number of people, but some percent of the people by right of their common properties. This is the same state with changing of place of work. For example the people, who changed their workplaces last years more times, they will likely change their workplaces this year too.

We simulate the reality well, if the person’s death is not dependent from how many examined people have died already, but his or her age, sex, habitat or financial problem.

Actually this is the main point of the microsimulation: the salary or the death of a person is independent from the other people, it depends on her/his own data.

II.2. Usage of microsimulation in the bank sector

In an other study the effect of the new products and processes of the banks are examined. There were four kind of data file in this project: demographic, consumption, income and bank data files. They were joined by Satistical Matching. In this complex data system the behaviour of the customers will be well examined. Data mining has two criterias. The first is: enough data and attributes about the customer and the second is the high level of quality. The bank has enough information about the customers, but it has problem with its quality.We know the product usage of the customers, but we do not know anything about the person. In the bank sector the demographic properties of the customers are difficult to determine, because the data protection law does not allow to get detailed data. The existing demographic data(name, address, ID number), which are asked by contracting are not actual in generality. The bank does not renew the data of the customers only if the customer has resort to a new service or product, where it has to give its data again. After some years the bank does not know anything about its customer. It knows only that the customer is a „good"(how often does it use ATM, does it pay its credit in time) or a „bad" customer. The other big problem is, that very slim information is stored. The marital status, school qualification and salary are not stored. The bank differentiate its customers only by right of sex, age and name.

The question can be arise why it is important to segmental the customers by demographic properties, because the purpose of the bank is only the profit and it is determined by the quality of the customer(good or bad and it pays or not). But if the purpose of the bank is to maximize the profit, the customers must be segmented by demographic properties. Think about the age! The frequency and the volume of the service used by a young or an active person are different from a pensioner's one. Different services are worth to offer for the different groups and with this activity we can amend the marketing. With the annex of the demographic and consumer missing data the quality of the marketing and the salesmanship can be amended. The main idea of the research was to estimate the demographic properties of the customers with the help of microsimulation and Statistical Matching of a nationwide representative existing data collection. To use these method we need at least two different tables. Either of the table contains the account information and the other one contains the demographic data of the customers. With the linking of the table the customers of

the bank get demographic properties. The main point of the method is to increase the number of the demographic attributes of the customers of the bank, so we can separate the customers by data mining technology and determine and evaluate the habits of the grouped customers.

With this method, we can decrease the high cost of the real testing and the risk of intruducing a new product.

The safety of the data is very important in the bank sector. The estimation on the real data is forbidden, so we have to create a research data file for the bank. The employees of the bank should learn this method to be able to use the microsimulation alone to avoid the occurrent data theft of the external developer or expert.

II.3. Microsimulation in telecommunication

The telephony habits of the customers are known, but the person is not known, because it is very difficult to determine the demographical properties of the people from the telephony habits. The data protection law and the business policy do not allow the detailed data measurement.

The existing demographical data(name, address, id number), which are asked from the customer, when the contract is signed are not actual yet. The supplier do not update these data, except for the customers, who enter into a new conract. (In this case, it asks for the data again.) After few years the supplier does not know anything about the customers. It knows only, that the customer is a good or a bad customer.(Pay its account in time or not.)

A big problem is, that the number of the demographical variables are too slim and they are not usable for analysis. The age, marital status, school qualification and the salary are not stored.

The most important problem is, that sometimes the party to the contract and the user of the service are not the same person. In this case the supplier does not know the user.

To be able to give the best marketing advertisement, the supplier has to know its customers. For example the students and the pensioners do not telephone as long as the managers. It is worth to separate the people. It offers telephone calls for the students and pensioners at a low price and it can use a higher price for the managers. There are a lot of packages of telephone services. You can choose among them: low sms cost, high call price or high sms cost and low call price. Of course if it would like to sell the first package, it has to look for the customers, who write sms very often and they call hardly.

To maximize the profit the customers must be separated by demographic properties. Think about the age! A joung people use the service in an other time interval, like the

pensioners. So it is worth to offer different packages for the different customers to amned the marketing.

Our purpose was to amend the quality of the marketing with the annex of missing demographical data. The idea was to estimate the demographical properties of the customers with Statistical Matching by right of an existing nationwide data file. We used two different tables. The first table contained the telephony habits and turnover data of the customers and the other contained the demographical properties.

With the linking of these tables the customers got demographical properties too, so we can separate them to be able to do a professional marketig.

II.4. Census in Serbia

We also have a study about how to substitute missing data with Statistical Matching. The „Vajdaság": a province in Serbia has not any detailed data about census, it does not contain financial data. Since a lot of Hungarian leave in Vajdaság, we substituted the missing data with the data of South-Hungary.

The cencus data for 2002 can be downloaded from the web page of Serbian Statistical Office. The following data are available: citizenship, age, sex, marital status, religion, school qualification, profession, but the financial data was not publicated. We would have liked to substitute these missing data. The donor data file was the data of South-Hungary. It is practical to use the data from that area, because it is very similar to Vajdaság, if we examine the professions, geographical properties, area of the places or the historical facts. If we assume, that the distributions(school qualification, size of areas, …) are the same, then the incomes must be almost the same.

To solve this problem we needed the census data of Hungary and Serbia too. Both of the databases have to be unweighted. If the unweighted data is available, we assign the most suitable South-Hungarian records to the Serbian ones with Statistical Matching. We got a complex table of Serbian data with income data.

III. Summary

In our opinion if the governmental organistions share their data with each other, they would get a complex data system. They should join their data files/tables, but they are unresolved. Everybody protects its data and use only for own usage. They should recognize, that sharing data would be useful for them and it is not a bad thing.

We offer everybody not to use real and expensive testing or measurement but also microsimulation, because you get almost the same result and it takes much less time to do it, than employ questioners to go to houses. Except that there are some cases, in which you can not ask peoples because of data protection law. We can modell their behaviour. We tried to explain the method of microsimulation in some words, but if you have further questions, please contact us.