A BOTTOM UP APPROACH TO ESTIMATE PRODUCTION-CONSUMPTION MATRICES FROM A SYNTHETIC FIRM POPULATION GENERATED BY ITERATIVE PROPORTIONAL UPDATING

Omar Abed, Tom Bellemans, Sungjin Cho, Davy Janssens, Geert Wets,
Transportation Research Institute [IMOB],Hasselt University, Wetenschapspark 5,
3590 Diepenbeek, Belgium
Gerrit Janssens
Group Logistics, Hasselt University, Wetenschapspark 5, 3590 Diepenbeek Belgium

ABSTRACT

In order to model freight movements in a region or nationwide, data about regional production and consumption of goods is necessary. This type of data is typically obtainable on an aggregate level only. The data takes the form of production - consumption (PC) matrices with individual cells representing the amount of goods being produced and consumed at the origin and destination respectively. The cell values of the PC matrices are in essence, an aggregation of individual firm to firm relations in form of goods production, processing and consumption. Iterative Proportional Updating (IPU) is a mathematical technique, which generates synthetic populations using a representative sample of the desired population. The use of this technique in transportation domains is typically to generate synthetic individuals-households populations, but by using a descriptive enough sample of firms, IPU can generate a synthetic firm population preserving real life firm attributes and distribution. The quality of the synthetic firm population relies greatly on the quality of the sample used. A main advantage of using this approach is that with a small sample of real firms with well-defined desired attributes, modellers are able to regenerate a total firm population. This means if firm production volumes can me estimated, firm to firm goods transactions can be reproduced on a nationwide scale, opening the door for a bottom up microscopic modeling approach of demand and supply relations. Another advantage of the IPU approach used in this paper is that, using a synthetic firm population enables us to evaluate the effect of policy impacts such as land use and firm location choice policies. This approach is also in agreement with the current direction in modeling freight movement at a microscopic level, as opposed to the traditional four steps based macroscopic approach. In this paper we present how to obtain regional PC tables from an IPU generated firm population, from which PC matrices are easily calculated. We then compare our findings with traditional way to obtain the PC tables and analyse results.

1. INTRODUCTION

Modelling freights movements on a disaggregate level has witnessed increasing efforts recently. One main reason behind such efforts is the wish to move from the traditional four steps modelling approach used to model private transport, towards a micro level simulation of goods movement [1,2,3]. Additionally, there is a need for such disaggregate data as input to a number of recently developed integrated transport and land use microsimulation models [4,5] and more recently [6]. This has become possible now thanks to gains in computing power, data availability and modelling approaches such as agent based modelling techniques [7]. Some agent based freight model frameworks as in [8] and its extension FREMIS [9] as well as operational models [3,10] exist today.

A pre-requisite to model freight movements, is the availability of regional input-output data on good movements, usually in units of tons per specific goods categories. Originated from Leontief's early input-output model in the 1930's [11,12] , these tabulations have witnessed many extensions during the years. One form commonly used and related to this work, takes the form of Production-Consumption (PC) matrices, which are basically a tabulation of the amount of goods per different categories produced in one geographical zone and consumed in another. Such tables are usually available only on an aggregate level [13,14]. This is mainly due to either, privacy concerns of firms, preventing opponents from obtaining competitive information, or due to the high cost of obtaining such data [3,15].

Synthetic population generation techniques provide a solution to this, by providing a synthetic yet reality compliant population of firms, distributed over the geographical area of interest. Iterative Proportional Updating (IPU) is one mathematical tool to do this and is used in this work. Extensively used for individual / households synthetic population problems , the use of the technique for synthetic firms population has received less attention see [16]. Another effort which combined modeling both individuals and business establishments was done by [17] . IPU uses in essence the complete and desirable attributes of a sample from the total population whose details are at hand, and re-weights it to obtain other members of the full population whose totals data only is available. IPU's main difference with its original form IPF (Iterative Proportional Fitting) , is that it is able to generate populations with interdependent dimensions. An example of such a population in this work is the number of firms, per size, per zone and corresponding industry types.

Synthetic populations offers modellers several advantages. First, it allows to overcome statistical errors and data incompleteness [18]. Till this date, in order to model annual freight movements for Flanders (northern part of Belgium) , one had to rely on government provided aggregate data. This data is basically, nationwide production, import and export volumes. Some disaggregation on a zonal level was then achieved using proxy metrics like population density and employment rate. The goods are then distributed over all zones in Flanders using a Gravity law formula to link production with consumption. Additionally, time consuming pre-processing steps were needed for each run, and to obtain futuristic scenarios, data was reweighed by varying the yearly Gross Domestic Product GDP index. Moreover, such a technique allows us - through marginal data manipulation - to sense effects of policy measures related to firms [18].

A population synthesizer based on IPU is used in this work, to obtain a synthetic population of the total producing firms distributed in Flanders, for three types of goods. The firms data available contains information on industry type and size of firm, being either very large, large, medium or small. The example industries being 1.food, beverages and tobacco, 2. Raw building materials, 3. Chemical and pharmaceutical. It will be shown that IPU is able to reproduce closely the real life production firms population, of which data is available at hand [29]. This approach has an important advantage, as it provides us with a more disaggregate level of information about what happens inside a zone. It offers us information on the number of firms per industry type per size category inside a specific zone. From this information and using available data on cost of tons per good category and yearly turn over for each firm sizes, we estimate a per firm production volume. When aggregated to the corresponding zones, those figures represent yearly zonal production. Then we derive consumption figures making the assumption that approximately 54% of what is produced in Flanders stays in Flanders, an observation obtained from the Planet model [19]. Finally, to obtain the final PC matrices a GIS step is applied where a Gravity law formula matches production with consumption. We will then discuss our findings and how they relate to similar values obtained using the traditional way. We then conclude by summarizing suggested ways to improve results and future work directions.

2. PROBLEM AND METHODOLOGY

In following sections we discuss the background of the problem at hand and the proposed algorithm used to solve it. We also describe the data available and used in the process.

2.1. Problem Background

The process of modelling freight movement across Flanders so far, involved time consuming pre- processing steps. The data available on regional PC volumes are on an aggregate level only, and the process of matching demand with supply was largely based on proxy values such as population density, employment rate and yearly GDP index. It is also difficult to evaluate the effect of policy application. Therefore, a more disaggregate and bottom up approach to estimate PC volumes, resulting from firms population producing and consuming goods was a needed. Using IPU and available data on real firm population enables us to overcome above mentioned limitations. Manipulating totals margins in the IPU process enables us to sense the effect of some policies related to land use and firm location choice.

2.2. IPU Algorithm

IPU (iterative proportion updating), proposed by Ye et al. in 2009, is a synthetic population technique that aims at predicting population by using a small complete sample and marginal data in a target. The basic algorithm behind IPU is based on IPF, which iteratively adjusts a sample contingency table to marginal totals in a target to compute a proportion of agents (e.g. household, person or firm) with the combination of considered attribute categories [27].

IPU has distinct methodological features from the IPF. First, IPU can adjust multiple levels of agents while IPF only deals with one level, for example either household or persons, at a time. Therefore, IPU is well suitable for predicting synthetic population which consists of more than one level, for example both household and person at the same time. Second, IPU uses a sparse list-based data structure which is composed of a list of several records representing agents, while IPF uses a contingency table where multi-dimensions of attributes are considered. The sparse list-based data structure has an advantage over a contingency table. For instance, when more attributes are dealt with, the size of a contingency table grows exponentially following increase in dimensions considered. On the other hand, the size of the sparse list-based data structure grows linearly due to an increase in

columns,. Thus, IPU is more suitable for synthesizing population with rather several attributes by means of efficient data storage.

An IPU run is normally divided into three steps, as following [21]:

- In a pre-processing step, put a weight into individual record with initial value of one.
- In a fitting step, compute individual weights by updating itself with a proportion of a total frequency to a total marginal for every attribute categories.
- In a drawing step, estimate synthetic population by draw the desired number of agents according to the computed agents' weights using Monte-Carlo (MC) sampling.

The fitting and drawing steps are continually run until a result from each step meets a given threshold value of an error measurement or an iteration number reaches a given top criterion. A standardized root mean square error (SRMSE) is used for the error measurement in this study. The SRMSE is calculated as follows [28]:

$$\mathrm{SRMSE} = \sqrt{\sum_i \sum_j \frac{(o_{ij} - e_{ij})^2}{m \times n}} \Big/ \left[\sum_i \sum_j \frac{o_{ij}}{m \times n}\right]$$

where e_{ij} is the estimated number of population elements with attributes i and j, and o_{ij} is an observed number of population. m and n are the number of attribute values for attributes i and j, respectively. A zero value of the SRMSE means 100% match, and '1' means no match between estimated and observed value.

2.3. Data Used

2.3.1. BELFIRST firm database

Bel-first is a firm database for Belgium and Luxemburg provided yearly by Bureau Van Dijk [29]. Firms can be filtered out using several criteria. Attributes of interest for us and used in this work are number of firms, industry classification, address information, size (C1: very large, C2: large, C3: medium or C4: small), yearly turn over for last year available and number of employees. As explained later, yearly turnover will have a direct impact on estimating firm production size.

2.3.2. PLANET Model

The PLANET model is a model of the Belgian Federal Planning Bureau that models the relationship between the Economy and Transport [19]. The aim of the model is to produce: (1) Medium and long-term projections of transport demand in Belgium, both for passenger and freight transport (2) Simulations of the effects of transport policy measures (3) Cost-benefit analyses of transport policy measures. Regional production and consumption estimates for different goods categories in tons obtained from this model are then distributed among zones using a gravity law formula. Monetary values of tons per good type used in this work, are part of input files to this model.

2.3.3. Zoning system

The present-day Flemish Region covers roughly 13,522 km2 and is divided into five provinces, 22 districts and 308 municipalities. Flanders is divided hence into 308 zones, each zone representing one municipality.

3. RESULTS

In the following sections we present results from IPU run, how to estimate regional PC volumes and matrices. We conclude by analysing findings and we briefly describe future related work.

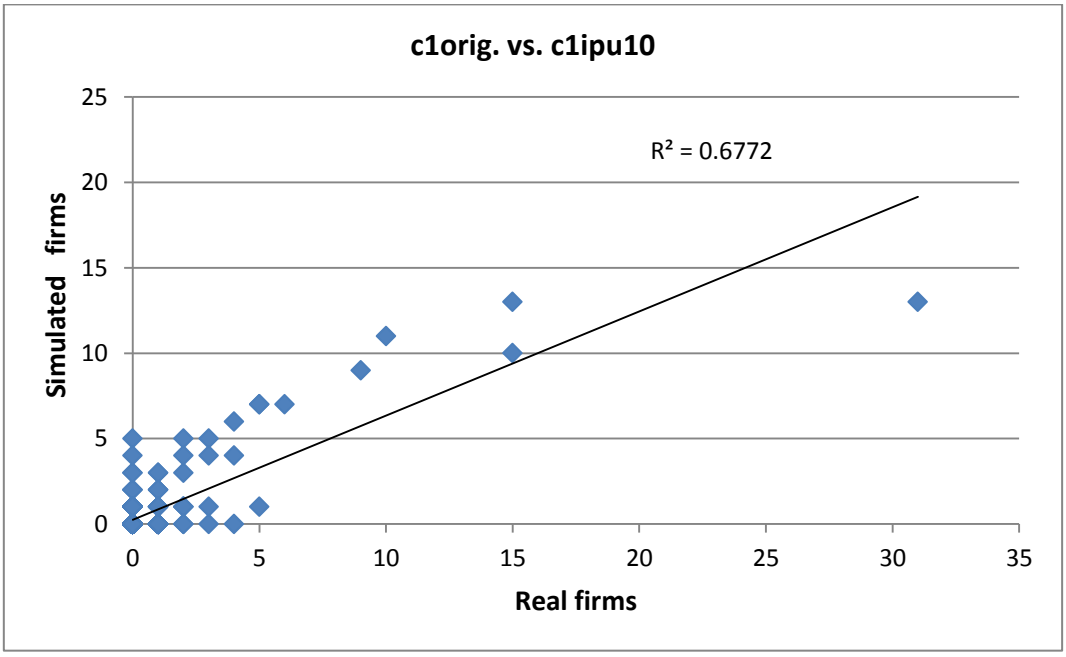
3.1. IPU Generated Firm Population

In our research study, the IPU process was applied for estimating synthetic companies in Flanders area. For that, we use 10% of a real population with complete information (in this study, four company sizes and three industry categories in all zones) as a sample for the fitting step. We then use this sample and marginal information to try to reproduce the full set (total of 6377 firms). *Bel-first* firm database is the source of the sample and marginal data [29]. For the drawing step, a MC simulation method was employed to draw the desired number of companies with a certain attribute category in a zone. Table 1 shows the performance of the IPU processing using a different size of sample. As seen in Table 1, there is no difference in the prediction accuracy, which is a Square Root of the Mean Square Error (SRMSE), in the fitting and drawing steps using a different sample size, but the processing time increases significantly.

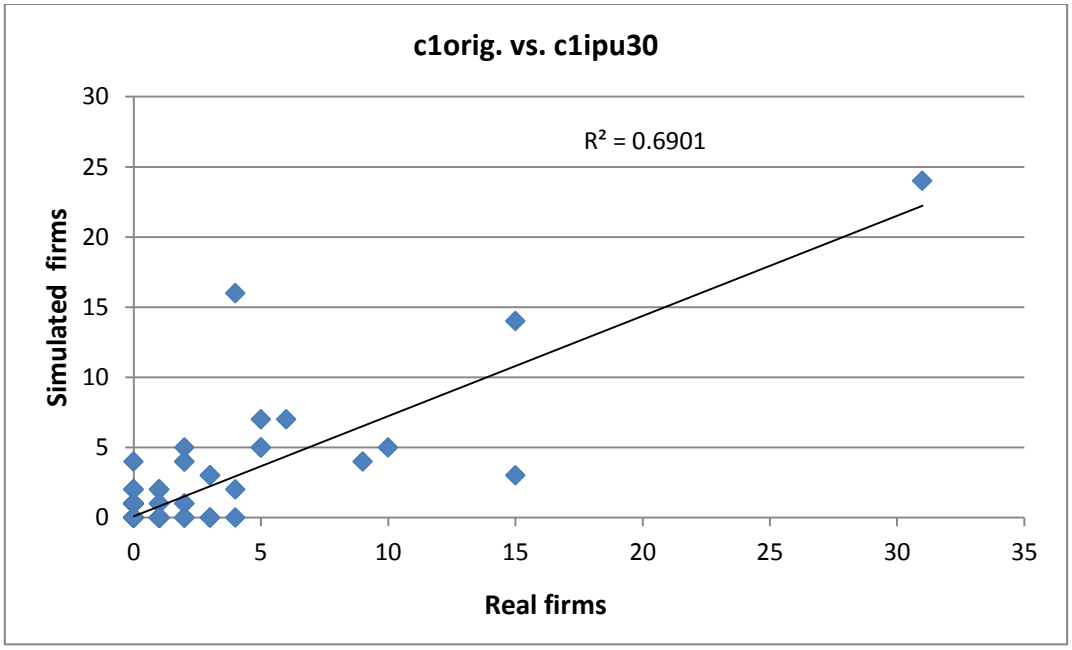
TABLE 1 Performance of IPU processing

Sample size	Error measurement (SRMSE)		Processing time
	Fitting	Drawing	
10% (722 firms)	0.0039	0.0345	51 sec.
30% (2176 firms)	0.0053	0.0353	7 min.
60% (4334 firms)	0.0034	0.0344	28 min.

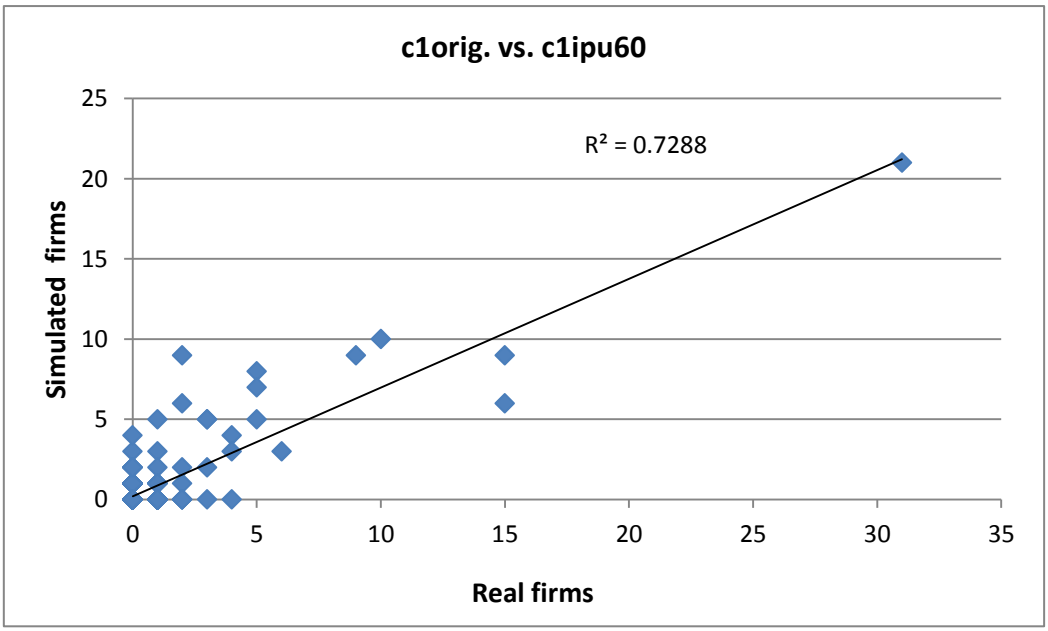
As previously explained, three sizes of sample data were used to see the effect of sample size on the quality of the reproduction of the whole population. In Figure 1 below we compare the real firm population of Very large firms (category c1) with respect to the synthetic one obtained for three sample sizes. It is clear that a larger initial sample size results in better population reproduction. Horizontal and vertical axis are number of firms from real and IPU generated respectively.



(a)



(b)



(c)

FIGURE 1 Number of c1 firms real population vs. IPU generated at (a) 10% (b) 30% (c) 60 % sample size.

Output data from IPU runs contain listing of total firms active in the three industries of interest, their zonal location and size. The next step will be to estimate per firm production size in tons for different goods categories. This will result in a more disaggregate level of information than previously

obtained, in the form of zonal totals for production and consumption. Using IPU output, combined with available data on yearly firm turnover and price of goods per ton, we can have a firm level production which when aggregated to a zone level will result in a bottom up estimation of zonal production values.

3.2. Regional PC Tables

Next, we obtained regional production and consumption tables. These are in essence production volumes in tons for the three goods categories, present in the 308 zones. To do this we will make use of available information on firm sizes and their corresponding yearly turn over in euros. The assumption made here is that every firm of the same size category will produce the same amount of goods within a good type. Averaging yearly turn over per firm size, and using data at hand for the cost in euros per ton for each good category, we obtain yearly production estimates per firm size, per good type. This is so far the production side. To estimate consumption volumes, we use the findings from Planet model that nearly 54% of what is produced in Flanders stays in Flanders. The remaining percentage is transported to other parts of Belgium and international trade. Using the synthetic firm population obtained from IPU step (60% sample size), and data on monetary values per ton for every category, Table 2 shows resulting production and consumption values for good type 2, in zone Antwerp as an example.

TABLE 2 Example PC calculations per firm size for good type 2, Antwerp.

Firm size category	Number of firms	Yearly turn over [Euros]	Production [Tons]
C1	2	144.684.265	4019007.36
C2	8	25.066.454	2785161.55
C3	28	4.244.970	1650821.66
C4	58	276.840	223010

This step is repeated till similar PC values were obtained for all 308 zones. Figure 2 below shows a sample of obtained regional PC estimates from above procedure compared to similar estimates using the traditional approach. The vertical axis is amount of goods in tons, the horizontal axis represents zone names.

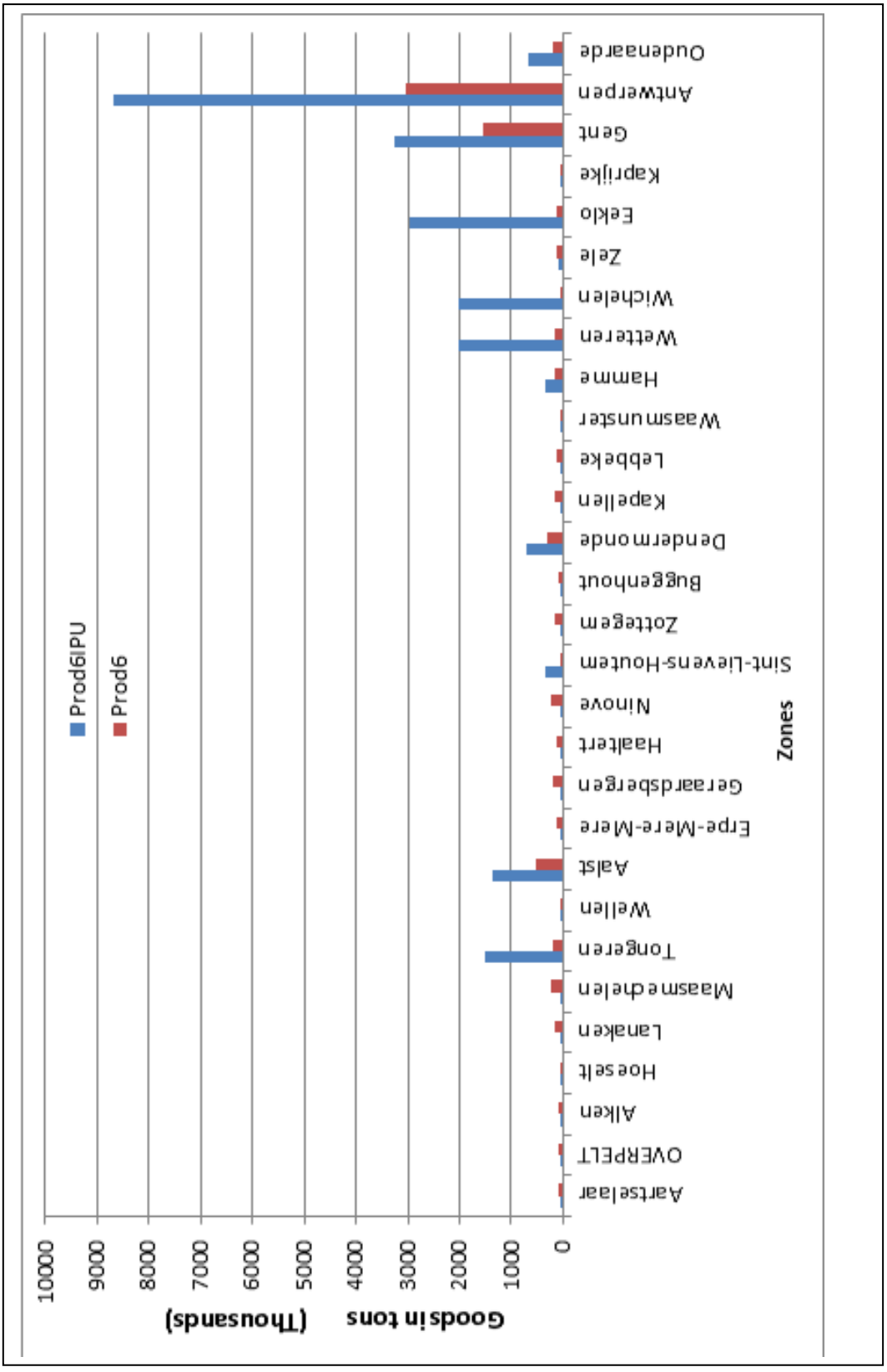


FIGURE 2 Regional production estimates from IPU population compared to traditional approach.

3.3. Generating PC Matrices

Now that we have generated regional PC pairs for every zone by aggregating PC firms volumes in each zone, building PC matrices is a straight forward task. This is in principle matching demand with supply. To do so, we simply apply a Gravity law formula in a GIS (TransCAD®) [30] processing step. The formula uses distance as impedance to good movements.

4. ANALYSIS

As we have shown, using IPU enabled us to build a synthetic firm population maintaining distribution characteristics of the real population. Furthermore, we were able to build PC estimates using firm attributes and monetary data. The obtained figures follow similar estimates obtained from using the traditional approach with a tendency to overestimate production volumes in zones with high firms density. This is mainly due to the turnover averaging effect we applied when calculating per firm size's production volumes. Several firms did not have any turnover reported. Moreover, for firms within a size category, the reported turnover values decrease exponentially, which means by averaging, several firms gets more than their real turn over. Additionally, some multinational firms report the total turnover value and not their corresponding one in Belgium.

5. CONCLUSION AND FUTURE WORK

In this research effort we showed that IPU can be to generate a synthetic firm population closely matching a real one in its size and distribution characteristics. Combined with real firm data available at hand, we were able to obtain PC tables from a more disaggregate level of information as previously done. In our approach we obtained PC volumes from firm level information on firm size, location, yearly turnover and estimated production volume. IPU has another advantage of enabling modellers to design future prediction scenarios and test some policy enforcement effects. A future line of work is to define firm to firm production-processing-consumption relations among different industries to have a more realistic estimation of PC volumes. The current approach does not take into account that some good types are processed or used as input into producing other good types. A better approach to estimate per firm production shares from information on yearly turnover will be investigated. A possible solution is to use information on number of employees to estimate turnover when no turnover is reported.

REFERENCES

[1] Wisetjindawat W., K. Sano, S. Matsumoto , P. Raothanachonkun. Micro-simulation model for modeling freight agents interactions in urban freight movement. In CD-ROM Proceedings, 86th Annual Meeting of the Transportation Research Board, Washington DC, January 2007.

[2] Samimi A., A. Mohammadian, K. Kawamura. Behavioural freight movement modeling, Presented at The 12th International Conference on Travel Behaviour Research, Jaipur, Rajasthan, India, 2009,13-18.

[3] Holmgren J., P. Davidsson, J. A. Persson, L. Ramstedt. TAPAS: A multi-agent-based model for simulation of transport chains. In *Simulation Modelling Practice and Theory*, Volume 23, 2012, Pages 1-18, ISSN 1569-190X.

[4] Moeckel R., K. Spiekermann, C. Schürmann, M. Wegener. Microsimulation of Land Use. In *International Journal of Urban Sciences* 7[1], 2003, 14-31.

[5] Miller E., J. Hunt, J. E. Abraham, P. A. Salvini. Microsimulating urban systems. In *Computers, Environment and Urban Systems* . 28 ,2004, 9–44.

[6] Ballas D., G. Clarke, D. Dorling, H. Eyre, B. Thomas and D. Rossiter. SimBritain: A Spatial Microsimulation Approach to Population Dynamics. In *Population, Space and Place* . 11, 2005,13–34.

[7] Harland K., A. Heppenstall, D. Smith and M. Birkin. Creating Realistic Synthetic Populations at Varying Spatial Scales: A Comparative Critique of Population Synthesis Techniques. In *Journal of Artificial Societies and Social Simulation*, 15, 1-1, 2005.

[8] Roorda M. , R. Cavalcante, S. McCabe, H. Kwan. A conceptual framework for agent-based modelling of logistics services. In *Transportation Research Part E* ,46 ,2010, 18–31.

[9] Roorda M. , R. Cavalcante. Freight Market Interactions Simulation [FREMIS]: An Agent-Based Modeling Framework. In *Procedia Computer Science*, 19, 2013 , 867 – 873.

[10] Liedtke G. Principles of micro-behavior commodity transport modeling. In *Transportation Research Part E: Logistics and Transportation Review*, Volume 45, Issue 5, September 2009, Pages 795-809, ISSN 1366-5545.

[11] Leontif W.. Quantitative Input and Output Relations in the Economic Systems of the United States. *The Review of Economics and Statistics*, Vol.18, No.3, Aug.,1936, pp. 105-125.

[12] Leontif W.. Interrelation of Prices, Output, Savings, and Investment. In *The Review of Economics and Statistics*, Vol. 19, No. 3, Aug. 1937, pp. 109-132.

[13] The Aggregate-Disaggregate-Aggregate [ADA] Freight Model System, Moshe Ben-Akiva, Massachusetts Institute of Technology and Significance, Gerard de Jong, Significance and ITS Leeds.

[14] De Jong G., M. Benakiva and J. Baak. Method Report -Logistics Model in the Norwegian National Freight Model System [Version 2]. Deliverable 6a for the working group for transport analysis in the Norwegian national transport plan. Project 07044, Dec. 2008.

[15] Müller K., K.W. Axhausen. Population synthesis for microsimulation: State of the art. Presented at Swiss Transport Research Conference, Switzerland,2010.

[16] Ryan J., H. Maoh, P. Kanaroglou. Population Synthesis: Comparing the Major Techniques Using a Small, Complete Population of Firms.In *Geographical Analysis* ,41, 2009, 181–203.

[17] Moeckel R., K. Spiekermann. M. Wegener. Creating a Synthetic Population. presented at the 8th International Conference on Computers in Urban Planning and Urban Management [CUPUM], Sendai, Japan, May 2003.

[18] Alfons A., S. Kraft, M. Templ, P. Filzmoser. Simulation of synthetic population data for household surveys with application to EU-SILC. Report within the 7th framework program for research Theme 8, Socio-Economic Sciences and Humanities, Project AMELI.

[19] Desmet R., B. Hertveldt, I. Mayeres, P. Mistiaen, S. Sissoko. The PLANET Model: Methodological Report. Report from the Federal Planning Bureau. April 2008.

[20] Lahr M. L., L. De Mesnard. Biproportional Techniques in Input–Output Analysis: Table Updating and Structural Analysis. *Economic Systems Research*, ISSN 0953-5314 print; 1469-5758,Vol. 16, No. 2, June 2004.

[21] Ye X., K. Konduri, P. Waddell, B. Sana, R. M. Pendyala. A methodology to match distributions of both household and person attributes in the generation of synthetic populations. Submitted for presentation at the 88th Annual Meeting of the Transportation Research Board January 11-15, 2009.

[22] Müller K., K.W. Axhausen. Hierarchical IPF: Generating a synthetic population for Switzerland. Presented at the European Regional Science Association, ERSA 2011.

[23] Edwards W. Deming and Frederick F. Stephan. On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals are Known. In *The Annals of Mathematical Statistics* Vol. 11, No. 4 [Dec., 1940], pp. 427-444.

[24] Csiszar I.. I-divergence geometry of probability distributions and minimization problems. In *The Annals of Probability*, Vol. 3, No. 1, 1975, 146-158.

[25] Fienberg, S. E.. An Iterative Procedure for Estimation in Contingency Tables. In *The Annals of Mathematical Statistics* 41 [3], 1970, 907–917.

[26] Pukelsheim F., B. Simeone. On the Iterative Proportional Fitting Procedure: Structure of Accumulation Points and L1-Error Analysis. Preprints - Herausgeber: Institut für Mathematik der Universität Augsburg [2009-05].

[27] Beckman, R. J., K. A. Baggerly, M. D. McKay. Creating synthetic baseline Populations. In *Transportation Research Part A: Policy and Practice*, 30 [6], 1996, 415-429.

[28] Pitfield, D. E.. Sub-optimality in freight distribution. In Transportation Research, 12 [6], 1978, 403-409.

[29] Bureau Van Dijk. Company information and business intelligence. *Bel-first* firm database for Belgium and Luxemburg. http://www.bvdinfo.com/Products/Company-Information/National/Bel-First. Accessed Jan. 10, 2013.

[30] Caliper Corporation**,**1172 Beacon Street, Suite 300 Newton MA 02461-9926, USA http://www.caliper.com.