*STATISTICS IN TRANSITION new series, December 2024*
*Vol. 25, No. 4, pp. 99–117, https://doi.org/10.59139/stattrans-2024-005*
Received – 28.02.2024; accepted – 02.08.2024

# AMUSE: Analysis of mobility using simultaneous equations. Present population of refugees in Poland

**Sebastian Wójcik**[1]

## Abstract

Due to the conflict in Ukraine, which escalated on $24^{th}$ February 2022, and caused a large inflow of Ukrainian citizens to Poland, a need to investigate this phenomenon by official statistics has arisen. When it comes to tracking the movement of refugees, statistical and administrative data sources fail due to the lack of timeliness or spatial granularity. Therefore, official statistics is reaching for big data sources which seem to be complementary to statistical and administrative data sources. In this paper, we deal with the synthetic Mobile Network Operator (MNO) daily data obtained from SIM cards issued to Ukrainian refugees by one of MNOs operating in Poland. We propose AMUSE, a workflow for data analysis, a model for the data deduplication and mobility estimation as well as a simple estimator of the present population. All these functions of AMUSE are based on the aggregated signaling data on time and territory.

**Key words:** mobility, Mobile Network Operator data, refugees, simultaneous equations, experimental statistics.

## 1. Introduction

The full-scale war taking place in Ukraine caused widespread damage of residential buildings, industrial facilities and critical infrastructure. Intense military operations led to mass migration of Ukrainian people, both inside the country and abroad. According to the UN Refugee Agency estimates, approximately 3.7 million people have been internally displaced and over 6.3 million emigrated abroad. Most people leaving Ukraine decided to cross the border into Poland. Admitting such a large number of refugees in a short time was a great challenge for the Polish authorities, non-governmental organizations and ordinary citizens. Some Ukrainians, after a short stay in Poland, went to other countries. However, many of them decided to stay in Poland for longer. This required creating a support system in the form of social benefits, health care and educational services (some recent reports concerning situation of Ukrainian can be found in References).

Large-scale inflow of refugees is a challenge for labor market, real estate market, education system, and health care system. For the process of planning humanitarian assistance and public services for refugees, information about the number of refugees being hosted becomes substantial. There are several approaches to estimating the number of refugees

[1]Institute of Mathematics, University of Rzeszów, Rzeszów, Poland. E-mail: swojcik@ur.edu.pl, & Statistical Office in Rzeszów, Poland. E-mail: s.wojcik@stat.gov.pl. ORCID: https://orcid.org/0000-0003-2425-9626.

including a sample survey, an administrative data source, Mobile Network Operator data, Payment Card Operator data, Social Media data, etc. Applied approach is heavily conditioned to the available data sources. The moment of crossing the border as well as applying for government services by a particular refugee should left a footprint in one or more administrative data sources. Thus, the administrative data sources, if available, seem to be a primary data source for estimating the number of refugees. It is not the case for many emergencies and crises across the globe, e.g. Afghanistan, Syria, and Sudan. It turns out that even in the case of countries with a well-developed system of registers, some other obstacles may occur in utilizing administrative data sources in estimating the population of refugees solely. Recently, Statistics Poland in close collaboration with the World Health Organization (WHO) carried out a sample survey of Ukrainian refugees in Poland (details can be found in the report *Health of refugees from Ukraine in Poland 2022. Household survey and behavioural insights research*). The survey was conducted by offices in Rzeszów and Lublin. The quantitative component was used to collect health information about the refugees while qualitative component was used to collect the experiences of refugees in accessing health services. To compute the estimates, the following administrative data sources were utilized: Border Guard data on evacuees of Ukrainian nationality at the Polish-Ukrainian border and population register (PESEL). It is worth noting that in the process of computing sample weights there was a substantial role of a pilot study conducted just after the war breakout which revealed that 54% of refugees decided to leave Poland and travel further to other countries. Since Poland is in the Schengen Zone, there is no regular border control within and so the movement of refugees from Poland to other countries in the Schengen Zone would be unregistered. Hence, without the pilot study, the number of refugees in Poland would be harshly overestimated.

The administrative data sources are often used to estimate the *usually resident population*, which is based on actual stay in the given area over a twelve months. Recently, some new concepts of population have been developed. One of them is the so-called *present population* also known as *de facto population* (Lanzieri (2013)). In opposition to the usually resident population, the present population is a snapshot, that is, it consists of all individuals present in a given area in the given moment of time. Some researchers (Lanzieri (2019)) advocate these alternative concepts of population as the complementary statistics providing information on population movements. The administrative data sources seem to be insufficient for the task of estimating the present population. Therefore, there is a need for an alternative data source. Letouzé and Jütting, representatives of the Data-Pop Alliance (a think-tank on Big Data and development) and Paris21 (The Partnership in Statistics for Development in the 21st Century), in their inaugural paper from 2015, argue that Big Data may provide faster and cheaper data with better granularity. Still, it shall be a complementary data source instead of being a replacement for standard surveys carried out by official statistics, including population statistics.

Mobile Network Operator data was utilized by official statistics, among others, in estimating the present population (Ahas et al. (2015), Deville et al. (2014)), mobility (Alexander et al. (2015), Diao et al. (2016)), and migration (Lai et al. (2019)). MNO data were also used to support policy against COVID-19 pandemic (Badr et al. (2019)). Nevertheless, there are research papers giving a warning of possible biases of population and migration

estimates based on MNO data (Wesołowski et al. (2013)) and e-mail data (Zagheni et al. (2012)).

Recently, Statistics Poland obtained Mobile Network Operator daily data pertaining to SIM cards issued to Ukrainian refugees by one of MNOs operating in Poland aggregated to LAU level. In this paper we deal with a synthetic version of data. The synthetic dataset preserves trends, seasonality and spatial dependency of the original dataset. The original dataset values were scaled and the noise was added. Hence, we propose a workflow for analyzing such type of a dataset in terms of seasonality and spatial dependency. In the main part of this paper, we propose AMUSE: a mobility model for the data deduplication and mobility estimation as well as a simple estimator of the present population.

## 2. Problem Statement

Statistical models of the present population are strongly embedded in the datasets obtained from MNO. These datasets include a cell plan and event data. The cell plan contains information about the geographical location of the Base Transceiver Stations (BTS) or alternatively *the cell towers* and their properties such as range, propagation direction, networks serviced, etc. The next figure presents the geographical location of the cell towers in the capital city (on the left) and in the small town, namely Ustrzyki Dolne (on the right).

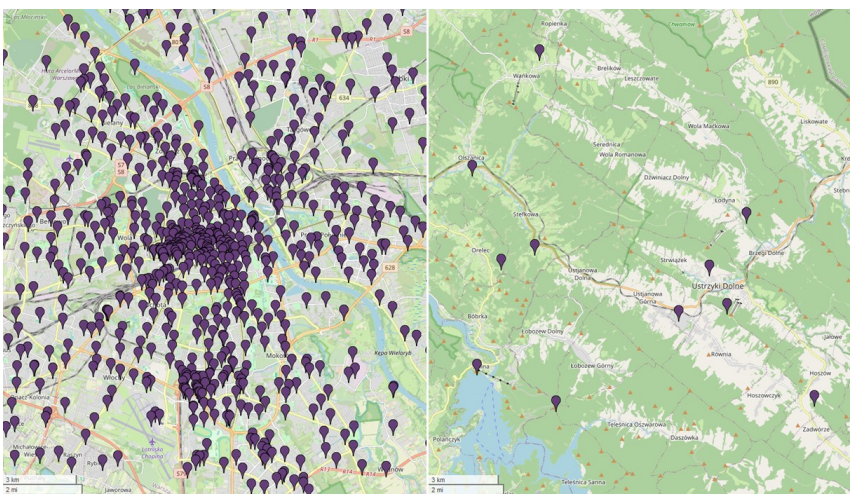


**Figure 1:** Cell plan

Density of BTS is really high in the capital city which has around 2 million population. On there other, in the small town with roughly 18 thousand inhabitants, there are only few BTS.

The event data contain information about the activity of particular cell phones. *Call Details Records* (CDR) consist of the records about calls (initiating and receiving) and SMSes (sending and receiving). Moreover, if CDR provide details on mobile data usage, then it is

often called *Data Details Records* (DDR) (Tennekes and Gootzen (2021)). Thus, CDR and DDR cover information about activity of the mobile phone users. MNO collects also *signaling data*, which are passive data consisting of logs to the cell towers. It is worth noting that CDR do not provide an information about the location of a particular mobile phone. The event record just states that in a given moment of time a particular mobile phone logged to a given BTS. Moreover, it is mostly a BTS with the strongest signal in a range of a particular mobile phone.

When dealing with MNO data, researchers and data analyst must keep in mind several issues concerning data quality aspects (Saidani, Bohnensteffen, and Hadam (2022)):

- Accuracy differences between MNOs - Variation in the events generated as a result of different practices between MNOs.
- Spatial accuracy - Short distances are underrepresented since location changes are only identified when a SIM card moves into a new mobile network cell. Moreover, smallest spatial unit varies greatly at the regional level due to different degrees of cell coverage, and accuracy of location estimation differs considerably.
- Asymmetric data losses - Units with little activity are disproportionately affected by anonymization losses.
- Validity - The assumption that a SIM card always communicates with the nearest antenna is not supported empirically.
- Biased features - Socio-demographic characteristics, e.g. age and gender are biased.
- Undercoverage - Not all sections of the population use a mobile device.

The event data may be provided by MNO in the form of micro data and aggregated data. In the case of micro data, one of the first steps includes modelling the spatial coverage patterns of BTS, that is mapping each cell tower to a geographical territory. This function is called *cell geolocation* (Salgado et al. (2020)). Ricciato et al. (2020) grouped existing geolocation methods into two families:

- Tessellations: after mapping geographical territories to cell towers, these areas remain disjoint. The simplest approach generates the simple fixed grid. A more data-driven approach includes Voronoi partitioning (Baccelli and Błaszczyszyn (2006)).
- Overlapping cells: mapping allows overlapping of geographical territories (Ricciato and Coluccia (2021)). Thus, it is a more general approach than tessellation.

The later step involves building a probabilistic model and estimating. In the recent literature several solutions can be found, namely:

- Maximum Likelihood Estimator based on multinomial distribution (Riccatio (2016)),
- Maximum Likelihood Estimator based on Poisson distribution (Shepp and Vardi (1982), Van der Laan, de Jongey (2019)),

- Simple Bayes-rule estimator (Tennekes and Gootzen (2021)).

In the case of aggregated data, the data can be obtained in various forms. In this paper we focus on the form of MNO data strictly connected with the use case of estimating the present population of Ukrainian refugees in Poland.

Several Mobile Network Operators issued SIM cards to Ukrainian citizens crossing the Polish-Ukrainian border after the war in Ukraine began. In order to register a SIM card in Poland for Ukrainian citizens, an identity document (ID card, passport or permanent residence card) is needed in the case of registration at registration points, or name, surname and PESEL (Universal Electronic System for Registration of the Population) in the case of registration through a bank. Due to the need to have an identity document or a PESEL number, the estimated results on the basis of Mobile Network Operator data should be comparable to the register of Ukrainian citizens who were assigned a PESEL number based on the Act published on March 12, 2022. Therefore, data obtained from a Mobile Network Operator generally does not include people who:

- came to Poland before the outbreak of the war,
- came to Poland after the outbreak of the war, but did not use a dedicated SIM card,
- registered a SIM card and then stopped using it (inactive card).

The obtained data are the daily counts of active SIM cards on Local Administrative Units level (LAU). We shall use also a term 'area' interchangeability with LAU. The MNO's system counts a given SIM card as active in a specific area provided the telephone with this card was active in that area for at least 3 hours a day. This means that if someone travelled between LAUs, SIM card could be counted multiple times in one day. Taking into account travel time between areas such a person could be assigned to a maximum of 7 neighboring or non-neighboring LAUs. Hence, we face a problem of multiple counts of particular SIM cards.

Let us note that the dataset does not contain any information on the directions of movements. Therefore, it is not possible to distinguish on its basis whether a given SIM card belonged to a resident of the given area or somebody who commuted to work, to school or to university as well as somebody who had a business or holiday trip. This distinction is essential to build a register of Ukrainian residents in Poland. Moreover, in the case of possibly longer business or holiday trip (at least three LAUs visited) we cannot distinguish which consecutive LAUs were visited within a trip. Deriving information how refugees commute and travel in Poland is also a subject of concern. Such information can be presented in the form of mobility matrix or series of mobility matrices. The idea of mobility matrix is incorporated in the proposed AMUSE model presented in details in the fourth section and it is a crucial tool in estimating the number of unique SIM cards.

After estimating the number of unique SIM cards from MNO, we need to estimate a total number of unique SIM cards from all MNOs operating in Poland. Finally, we shall derive a population of unique SIM card holders and then the whole population, that is, including persons who are not SIM card holders, especially children or some elderly persons. To this end, we start with data analysis for some insights.

## 3. Data analysis

The dataset contains information on the total daily number of active SIM cards issued by one of MNOs to Ukrainian citizens in the period from July to December, 2023 for 2,477 LAUs. No additional variables and no metadata are available. Since the data are already aggregated, there is no possibility to control quality issues on micro level. Further, we deal with issues of biased features and undercoverage.

The next plot presents the data aggregated to the national level. In the given period, the total number of active SIM cards increased almost by 10% (comparing the end points of the time interval). Strong deviation from the trend is observed, e.g. in September when the total number of active SIM cards decreased by 17% between 14 and 17 of September and then increased by 18% in a one day.

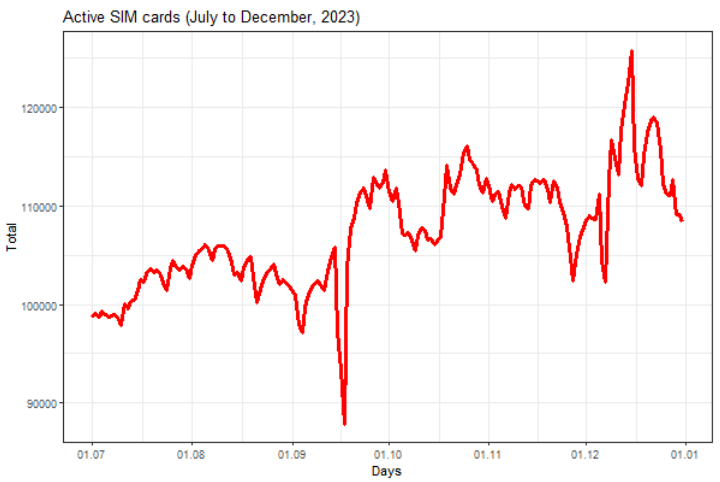


**Figure 2:** Active SIM cards time-series

The time-series analysis is ended with the seasonal decomposition. Seasonal decomposition by LOESS method proposed by Cleveland et al. (1990) was used. Figure 3 presents the results.

In the analyzed period, the rising trend is observed but with some turbulences. The seasonal part ranged from -1.6 thousand to 0.9 thousand while the remainder component ranged from -10.3 thousand to 6.3 thousand Weekly seasonality is not clearly visible. To investigate it, we used three seasonality tests, that is: Test for Seasonal Unit Roots proposed by Osborn et al. (1988), Test for Seasonal Stability by Canova, Hansen (1995) and Test for Seasonal Unit Roots presented in Hylleberg et al. (1990). Trigonometric version of Canova and Hansen test for seasonal stability indicated no seasonal cycles of seasonal frequencies $\frac{2\pi}{7}$, $\frac{4\pi}{7}$ and $\frac{6\pi}{7}$. Canova and Hansen test with dummy variables for seasonal stability revealed that on Friday there is significant and above-average traffic while on weekends the traffic is below the average. Osborn, Chui, Smith, Birchenhall test as well as Hylleberg, Engle, Granger, Yoo test indicated no significant seasonality.

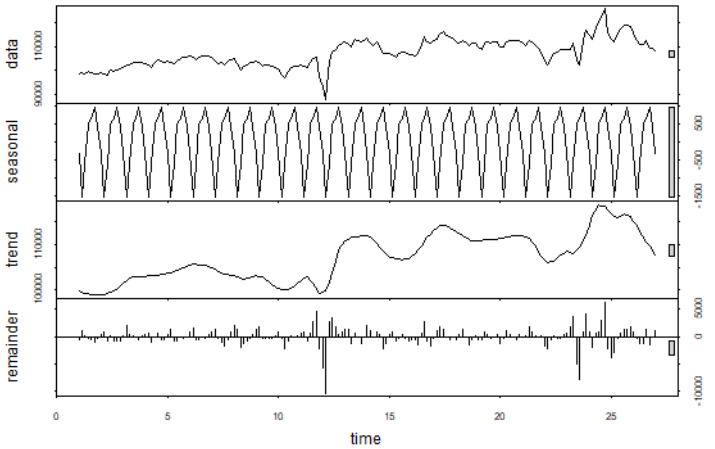


**Figure 3:** Seasonal decomposition of active SIM cards time-series

Let us investigate a spatial distribution of active SIM cards. The highest number of active SIM cards, for the majority of the analyzed period, was observed in the capital city of Warsaw. At the same time, there were 99 LAUs without a single active SIM card. The next figure presents the number of active SIM cards aggregated from LAU level to poviats (using former names, that is before 2017, aggregated from LAU level 2 to LAU level 1).

SIM cards mostly occur in urban areas, especially in big cities, or in tourist areas. One may ask if it follows a spatial distribution of population of the Polish citizens or the urban areas are in more favor of the refugees than rural areas. It may stem from the better access to accommodation, labor market, education and health services etc. To verify if this phenomenon occurs, let us recall first some studies on urban population. Auerbach (1913) observed a statistical regularity in the urban population of Germany. He discovered that the population of the city $P$ is proportional to reciprocal of the population rank $r$ in a decreasing order. Later, American linguist George Zipf (Zipf 1949) discovered that the similar regularity holds for rankings of words in text corpus for many natural languages. Zipf's law - applied to the population of the cities in a given country - states that the population of a given city follows the power law

$$P \sim \frac{1}{r^v} \tag{1}$$

where $v$ is a fixed parameter ($v = 1$ for Auerbach discovery). Zipf also proposed a rank plot, also called Zipf plot, which is used to investigate if Zipf's law holds. The next figure presents the Zipf plots for MNO data and LAUs' population in Poland.

The solid lines represent fitted values from the model given by (1). In both models, R-squared reached 0.99. Note that the coefficient $v$ amounted to 1.085 in the case of MNO data while it attained a value of 0.785 for Polish LAUs. Hence, SIM cards revealed that the Ukrainian refugees more often reside in highly urbanized areas, especially large cities. It could result from improved access to housing, labor market, education, health services, and so forth. Moreover, large cities are well connected in terms of rail and road transportation.

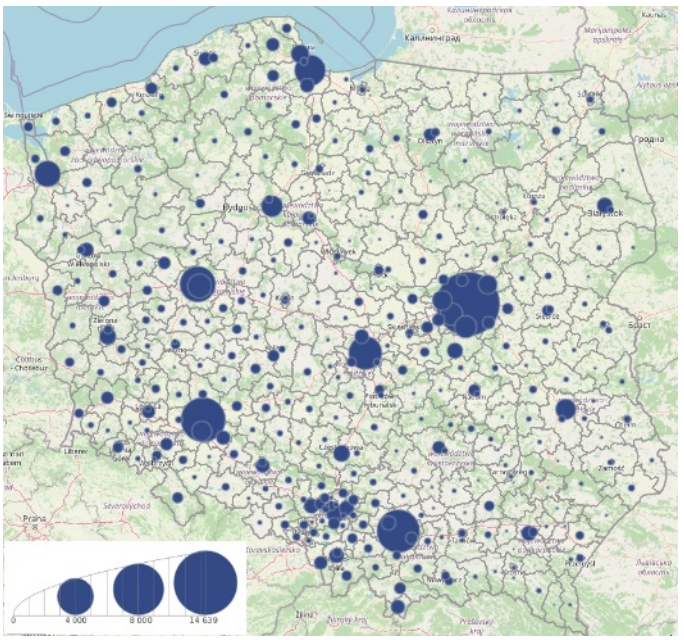


**Figure 4:** Spatial distribution of active SIM cards

## 4. AMUSE model

Due to the fact that a single SIM card could be counted several times, there was a need to build a model to estimate the number of unique SIM cards. Another important issue was to determine the mobility patterns of SIM card holders within the country. The following notations were used for modelling:

- $i \in \{1,...,n\}$ denotes index of spatial unit (area);
- $y = (y_1,\ldots,y_i,\ldots,y_n)$ denotes the vector of active SIM cards, $y_i > 0$ for $i \in \{1,...,n\}$;
- $x = (x_1,\ldots,x_i,\ldots,x_n)$ denotes the vector of unique SIM cards, $x_i \geq 0$ for $i \in \{1,...,n\}$;
- $T$ - number of unique SIM cards in total.

In the model, $y$ and $T$ are known, while $x$ is unknown. Determining $x$ is the goal of the first stage of the estimation process. SIM card holders may exhibit many different behaviors in terms of inter-area movements in terms of the number of areas visited, their neighborhood, etc. Intuitively, the most cases will be:

- single counting of SIM card holders who live, work, or study within the same area;
- double counting of SIM card holders who live in one area and work or study in another area. At the same time, we can expect more cases of double counting with neighboring than with non-neighboring areas.

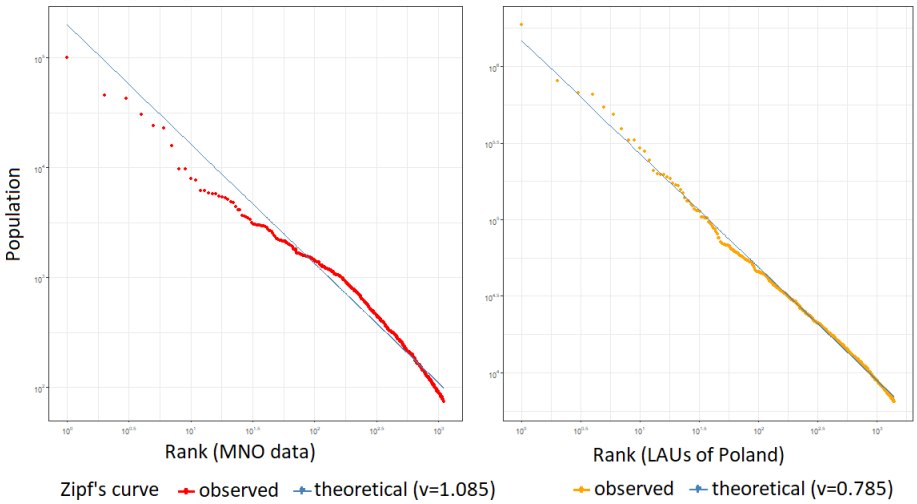


**Figure 5:** Zipf plots for MNO data and LAUs' population in Poland

In addition to commuting to the place of work or study, there are also movements related to trips for tourist and business purposes. These trips may involve more than two visited areas within a single trip. However, the number of such trips should be small when compared to all movement patterns.

### 4.1. Model setup

In order to build the model we start with the simplest case of a universe of two areas. Let us denote by $p_{12}$ a share of SIM card holders residing in the area number 1 who were counted also in the area number 2 and by $p_{21}$ a share of SIM card holders residing in the area number 2 who were counted also in the area number 1. Thus, by definition, $p_{12}, p_{21} \in [0,1]$, hence $x_i \leq y_i$ for $i \in \{1,2\}$. The next figure presents inter-area flows in the two-area universe.

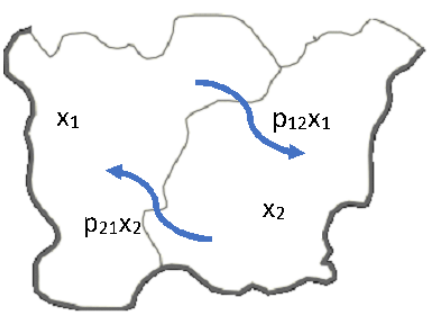


**Figure 6:** Flows in two-area universe

The number of active SIM cards $y_1$ in the area 1 is a sum of active SIM cards of residents $x_1$ and the number of active SIM cards of residents of the area 2, who visited the area 1 for

at least three hours, that is long enough to be counted by system of MNO as an active card. Therefore, the universe of two areas must meet the following equations:

$$\begin{cases} y_1 = x_1 + p_{21}x_2, \\ y_2 = x_2 + p_{12}x_1, \\ T = x_1 + x_2. \end{cases} \tag{2}$$

The system (2) of three equations contains four unknown variables. Hence, the system is undetermined.

In the case of a universe of, e.g. three areas, a particular SIM card can be counted even three times. For instance, resident of the area 2 could visit the area 3, and then could visit the area 1 or could visit the area 1 first, and then could visit the area 3. Hence, a more general model can be build two-fold:

- with dynamic approach,
- with static approach.

In dynamic approach we need to take into consideration a route followed by a particular SIM card holder. By route we understand a sequence of areas in which a SIM card holder were counted in a given day. Formally, let $s \in \{1,...,S\}$. Then, the $S$-step route is a finite sequence $(g_1,...,g_S) \in \{1,...,n\}^S$, that is a sequence of the area indices visited consecutively by a SIM card holder. When a resident of the area $g_1$ stayed within in a given day, then the route reduces to $(g_1)$.

Further, we can define the flow frequencies between areas taking into account a stage of a route. Let $s \geq 2$ and $(g_1,...,g_s,...,g_S)$ be a route. By $p^{(s)}_{ij|g_1,\ldots,g_{(s-1)}}$ we denote a share of SIM card holders who travelled from the area number $i$ to the area number $j$ in a $s$-th step of the route residing in the area $g_1$ after visiting the areas $g_2,\ldots,g_{(s-1)}$ (note that the visit must take at least three hours in our setting). Formally, $p^{(s)}_{ij|g_1,\ldots,g_{(s-1)}} \in [0,1]$ and $p^{(s)}_{ij|g_1,\ldots,g_{(s-1)}} = 0$ for $i = j$.

Let us consider a case of three areas. The next figure presents possible routes of travelling to the area 1 (on the left, routes from the area 3; on the right, route from the area 2).

Therefore, the universe of three areas must meet the following equations:

$$\begin{cases} y_1 = x_1 + p_{21}x_2 + p_{31}x_3 + p_{23}p^{(2)}_{31|2}x_2 + p_{32}p^{(2)}_{21|3}x_3, \\ y_2 = x_2 + p_{12}x_1 + p_{32}x_3 + p_{13}p^{(2)}_{32|1}x_1 + p_{31}p^{(2)}_{12|3}x_3, \\ y_3 = x_3 + p_{13}x_1 + p_{23}x_2 + p_{12}p^{(2)}_{23|1}x_1 + p_{21}p^{(2)}_{13|2}x_2, \\ T = x_1 + x_2 + x_3. \end{cases} \tag{3}$$

Compared to (2), the system (3) of four equations contains 15 unknown variables. In a general case, the number of unknown variables increases proportionally to $n^2$.

The dynamic approach would be preferred to static approach whenever detailed information about mobility is of interest. When it comes to the estimation of the present population,

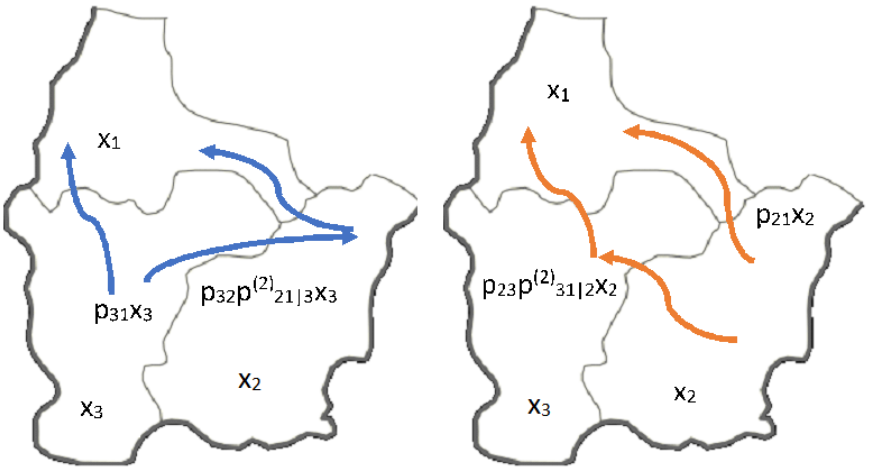


**Figure 7:** Flows in three-area universe

the static approach would be more straightforward and explainable. In the static approach we are only interested in the fact if a given SIM card holder visited a particular area. For instance, we are indifferent if a resident of $i$-th area visited $j$-th area directly or through the $k$-th area. That is, the routes $(ij)$ and $(ikj)$ are indistinguishable. In such a setting, the static model can be easily derived from the dynamic model. First note that we can rewrite (3) in the following way:

$$\begin{cases} y_1 = x_1 + \left(p_{21} + p_{23}p^{(2)}_{31|2}\right)x_2 + \left(p_{31} + p_{32}p^{(2)}_{21|3}\right)x_3, \\ y_2 = x_2 + \left(p_{12} + p_{13}p^{(2)}_{32|1}\right)x_1 + \left(p_{32} + p_{31}p^{(2)}_{12|3}\right)x_3, \\ y_3 = x_3 + \left(p_{13} + p_{12}p^{(2)}_{23|1}\right)x_1 + \left(p_{23} + p_{21}p^{(2)}_{13|2}\right)x_2, \\ T = x_1 + x_2 + x_3. \end{cases} \tag{4}$$

Putting $q_{ij} := p_{ji} + p_{jk}p^{(2)}_{ki|j}$ for $i, j, k = 1, ..., 3$, $i \neq j \neq k$, (4) takes the following form

$$\begin{cases} y_1 = x_1 + q_{21}x_2 + p_{31}x_3, \\ y_2 = x_2 + q_{12}x_1 + q_{32}x_3, \\ y_3 = x_3 + q_{13}x_1 + q_{23}x_2, \\ T = x_1 + x_2 + x_3. \end{cases} \tag{5}$$

Let $Q'$ denote transposition of the matrix $Q$ and assume that $1_n$ is a column vector of length $n$. In general, that is for an arbitrary $n \in \mathbb{N}$, the system of equations (5) can be presented in a matrix form

$$\begin{cases} y = Q'x, \\ T = 1'_n x. \end{cases} \tag{6}$$

The static approach has definitely less unknown parameters to estimate then the dynamic

approach. And so, the results are more data-driven than in the dynamic approach. On the other hand, it cannot be used to provide complex results on mobility. Nevertheless, it still captures a major part of movement, that is, commuting to the nearest area for the purpose of working or studying.

### 4.2. Priors

Since we are dealing with undetermined system of equations, the unknown $x$ and $Q$ can be estimated in an iterative way starting from a given prior $x_0$ and $Q_0$. The simplest form of the initial values $(x_0, Q_0)$ may be based on a single point in time, that is, without concerning changes of $y_i$ over time and any additional variables characterizing areas in terms of labor market, housing market, etc. Such a naïve prior may take a form

$$\begin{cases} \lambda = \frac{\sum_i^n y_i}{T}, & \\ x_i = \frac{y_i}{\lambda} & \text{for } i = 1,..,n, \\ q_{ij} = \frac{\lambda - 1}{n-1} & \text{for } i,j = 1,..,n,\ i \neq j. \end{cases} \tag{7}$$

Simple calculations show that the prior (7) satisfies $\sum_i^n x_i = T$. Moreover, $x_i \geq 0$ and $q_{ij} \in [0,1]$. Indeed, since $\sum_i^n y_i \geq T$, we have

$$0 = \frac{\frac{T}{T} - 1}{n-1} \leq \frac{\frac{\sum_i^n y_i}{T} - 1}{n-1} = q_{ij} \leq \frac{\frac{nT}{T} - 1}{n-1} = 1.$$

Further, (7) can be modified to take into account that $q_{ij}$ is positive, e.g. only for neighbouring areas. Then, denoting by $n_i$ the number neighboring areas of $i$-th area, we have

$$\begin{cases} \lambda = \frac{\sum_i^n y_i}{T}, & \\ x_i = \frac{y_i}{\lambda} & \text{for } i = 1,..,n, \\ q_{ij} = \frac{\lambda - 1}{n_i} & \text{for } i,j = 1,..,n,\ i \neq j. \end{cases} \tag{8}$$

The prior (8) is well-defined whenever $\sum_i^n x_i \leq 2T$ since that inequality ensures it holds $q_{ij} \in [0,1]$.

### 4.3. Estimation of the parameters

Now, we shall propose two methods for estimating the parameters of (6). The first one is based on the fixed-point iterations method. Among numerical methods for solving nonlinear equations or optimization problems, fixed-point iterations are widely used (cf. Shams et. al. (2022), Zhu et al. (2023)). These methods iteratively update an initial guess until a fixed point is reached, where the updated value equals the value from previous iteration. Their underlying idea share similarities with Banach's fixed-point theorem. Both involve finding points where certain conditions are satisfied, leading to estimation or convergence. The general idea is to define an operator $W : X \to X$ on a given space $X$ that transforms and updates a prior value $z_0 \in X$, that is $W(z_0) = z_1$. Repeating that transformation consecutively

$W(z_t) = z_{t+1}$ should lead to the solution $z^*$ satisfying

$$\lim_{t\to\infty} W(z_t) = z^*.$$

The solution is a fixed-point of the operator $W$, that is, it holds that $W(z^*) = z^*$. Existence of such fixed-point as well as convergence of the operator are conditioned to the properties of the operator $W$ as well as the properties of the space $X$ itself. Keeping that in mind we propose a fixed-point iterations method allowing updates of the values from the previous iteration only if the new values are in the domain of interest. For the learning rate $\alpha \in (0,1]$, the algorithm goes as follows:

(1) for $i = 1,...,n$ calculate

$$x_i^* = (1-\alpha)x_i + \alpha\left(y_i - \sum_{j,j\neq i}^{n} q_{ji}x_j\right)$$

and replace $x_i$ by $x_i^*$ in the $(i+1)$-th equation, provided $0 \leq x_i^* \leq y_i$. Elsewhere, skip the iteration.

(2) for $i, j = 1,...,n$, $i \neq j$ calculate

$$q_{ij}^* = (1-\alpha)q_{ij} + \alpha\left(\frac{y_j - \sum_{j,j\neq i}^{n} q_{ji}x_j}{x_i}\right)$$

and replace $q_{ij}$ by $q_{ij}^*$, provided $0 \leq q_{ij}^* \leq 1$. Elsewhere, skip the iteration.

In this procedure, $x_i^*$ is a convex combination of the current estimate $x_i$ and the estimate $y_i - \sum_{j,j\neq i}^{n} q_{ji}x_j$ satisfying the $i$-th equation. The same idea holds for $q_{ij}^*$. The learning rate controls the speed of convergence to the set of values satisfying (6). For the prior which is not a data-driven, e.g. of the form (7), it is advised to use a small learning rate which shall produce estimates within a range of valid values more likely. On the other hand, keep in mind that the small learning rate will increase a computational burden.

The second method is based on the optimization approach. The optimization criterion can combine the relative squared change or relative squared error for $x_i^*$ and Kullback–Leibler divergence for $q_{ij}^*$. Thus, the problem is to minimize

$$L(x^*, Q^*) = \sum_{i=1}^{n}\left(\frac{x_i^*}{x_i} - 1\right)^2 + B\sum_{i=1}^{n}\sum_{j,j\neq i}^{n} q_{ji}\log\left(\frac{q_{ji}}{q_{ji}^*}\right) \tag{9}$$

subject to condition (6) for $x^*, Q^*$. The parameter $B$ in (9) serves to set a trade-off between emphasis on relative squared change and emphasis on Kullback–Leibler divergence. Moreover, for Kullback–Leibler divergence we adopt a standard convention that $0\log 0 = 0$. The optimization approach is harder to implement but it has sound statistical background.

## 5. Estimation of the size of the refugees' population

In this chapter, we present a simple approach to estimating the size of the refugees' population. Let us recall that after determining the AMUSE model developed in the previous step, the number of unique SIM cards for one analyzed MNO was derived. Hence, the next step is to compute the number of unique SIM cards for all MNOs operating in Poland. Next, having information about the total number of unique active SIM cards, we shall proceed to determine the size of the refugees' population.

The basis for computing the estimates were reports published by the Office of Electronic Communications. The reports contain information on, among other things, the number of users, mobile traffic, revenues, market shares, and types of services provided. Some data are presented in additional breakdowns, e.g. divided into SIM cards and M2M, pre-paid, post-paid, by operators, etc. The reports include data on several services including mobile telephony service, Internet access service, VoIP telephony service, landline telephony service, bundled services, and paid TV services. Various services have different level of market penetration. According to the state as of the end of 2022 in the telecommunication services market (based on the aforementioned report), there were 52.6 million SIM cards, 6.7 million M2M SIM cards, 17.91 million Internet users, 13.92 million subscribers to bundled services, and 10.83 million subscribers to paid TV services. Among the services provided by operators, mobile telephony services have by far the widest reach. For this reason, operator shares in mobile telephony services were taken into account in further calculations. Table 1 presents market shares in 2022.

**Table 1:** Market shares of MNOs in 2022

| *MNO* | *market share* |
|---|---|
| P4 | 30.2 |
| Orange | 26.6 |
| Polkomtel | 20.4 |
| T-Mobile | 19.2 |
| Others | 3.6 |

Due to the very limited scope of available data on the mobile network market concerning citizens of Ukraine, in particular the absence of such data at the Office of Electronic Communications, it was necessary to make a series of assumptions:

(i) The structure of operators shares for all SIM cards in the market is similar to the structure of SIM cards issued to Ukrainian refugees.

(ii) The spatial distribution of SIM cards issued to citizens of Ukraine is similar for each MNO.

(iii) Movement patterns are similar for each MNO.

The first assumption could be partially verified based on online sources referring to the number of SIM cards issued to citizens of Ukraine. The information pertained (depending on the source) to only two or three operators. From the most recent data, which only covered

two operators, it was found that by March 16, 2022, they had issued 275,000 (57.9% of the total number of SIM cards issued by both operators combined) and 200,000 (42.1%) SIM cards, respectively. It turns out that the market shares of mobile telephony services of these two operators are very similar and remain in a proportion of 58.1% to 41.9%.

Let us denote the total number of SIM cards issued by $k$-th MNO to Ukrainian refugees and to Polish citizens by $S_k^{UA}$ and $S_k^{PL}$, respectively. Taking into account the assumption $(i)$, the estimator of the number of SIM cards issued to Ukrainian refugees by all Polish MNOs denoted by $S^{UA}$ can be given by

$$S^{UA} = S_k^{UA} \cdot \frac{\sum_k S_k^{PL}}{S_k^{PL}}. \tag{10}$$

Observe that (10) can be interpreted as a direct estimator (Horvitz-Thompson estimator) of the total population provided a simple random sampling without replacement with inclusion probabilities equal to $p = \frac{S_k^{PL}}{\sum_k S_k^{PL}}$. Then, under the assumption $(ii)$, the estimator of the number of SIM cards in $i$-th area $S_i^{UA}$ can be computed in the following way:

$$S_i^{UA} = S^{UA} \cdot \frac{x_i}{\sum_i x_i}. \tag{11}$$

It should be borne in mind that, according to the law, each SIM card must be registered and assigned to a person or company. In particular, multiple SIM cards can be registered to one person. In recent years, their number has been around 50 million in Poland, which gives an average of over 1.32 SIM card per person. On the other hand, individuals aged 13 or older are eligible to register a card. Consequently, due to legal and other circumstances (e.g. the level of digital literacy among different age groups), not every person owns a mobile phone and a SIM card. According to a survey conducted by the Office of Electronic Communications, 78.0% of Polish people (at age 15 or more) own a smartphone.

By $N$, $N_{SIM}$ and $N_{holders}$ let us denote the size of population, the total number of SIM cards and the total number of SIM cards holders, respectively. Then, note that we have the following equality:

$$N = \frac{N}{N_{holders}} \cdot \frac{N_{holders}}{N_{SIM}} \cdot N_{SIM}. \tag{12}$$

Moreover, $P := \frac{N_{holders}}{N}$ can be interpreted as a percentage of persons with SIM cards while $M := \frac{N_{SIM}}{N_{holders}}$ gives an average number of SIM cards per person. In a result, we obtain

$$N = \frac{N_{SIM}}{P \cdot M} \tag{13}$$

The equality (12) holds when all indicators pertain to the same market. While estimating the number of Ukrainian refugees on the basis of SIM cards, two statistics, that is *percentage of persons with SIM cards* and *average number of SIM cards per person*, are not known for this population and are replaced by the corresponding statistics from the Polish market.

In the case when age-gender characteristics of refugees differs from age-gender characteristics of the host country, the indicators *percentage of persons with SIM cards* and

*average number of SIM cards per person* can be harshly biased. If the age-gender cohort structure of refugees is available from, e.g. sample survey, then it can be used to weight the aforementioned indicators with respect to cohorts. It should be kept in view that age-gender characteristics of Ukrainian refugees in Poland can be investigated through, e.g. the register of Ukrainian residents under temporary protection, which was developed due to the act *Council Directive 2001/55/EC of 20 July 2001 on minimum standards for giving temporary protection in the event of a mass influx of displaced persons and on measures promoting a balance of efforts between Member States in receiving such persons and bearing the consequences thereof* and Polish act *Ustawa z dnia 12 marca 2022 r. o pomocy obywatelom Ukrainy w związku z konfliktem zbrojnym na terytorium tego państwa (Dz.U. z 2023r. poz. 103 z późn.zm.)*. Enormous disparities between age-gender cohorts of Polish citizens and Ukrainian refugees are presented in the next figure.

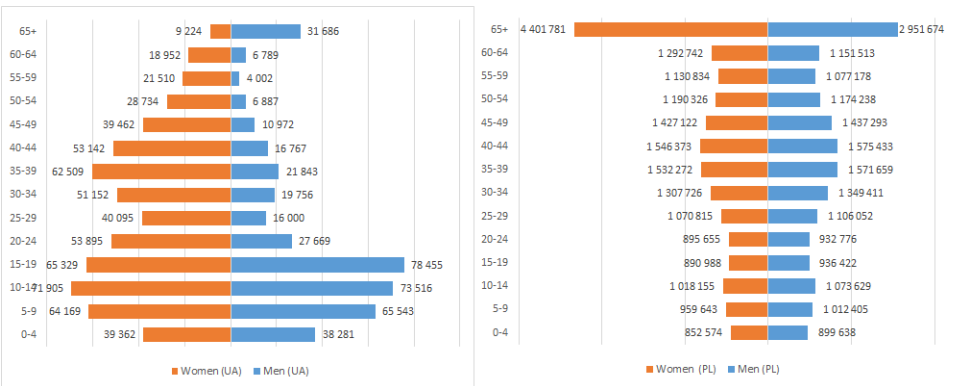


**Figure 8:** Age-gender cohorts of Polish citizens and Ukrainian refugees (as of May 31, 2023)

Size of cohorts was derived from the register of Ukrainian residents under temporary protection and Polish population register. Consider that the population of refugees mostly consists of children and young females. Adult (but not retired) males or elderly females are in the minority. Note that most of the children do not own smartphones formally. Since there are a lot of children in the population of Ukrainian refugees, the percentage of persons with SIM cards is lower than for the Ukrainian refugees in general.

Applying (10), (13), taking into account data from the Office of Electronic Communications and age-cohorts statistics from the register of Ukrainian residents under temporary protection, we obtained that for, e.g. MNO P4, to estimate the total number of Ukrainian refugees in Poland, the total number of unique SIM cards should be multiplied by 5.087362.

## 6. Conclusions

The article discusses the significant influx of Ukrainian refugees into Poland following the escalation of the conflict in Ukraine in February 2022. It highlights the challenges in tracking refugee movements using traditional statistical and administrative data sources

due to issues such as timeliness and spatial granularity. As a result, official statistics are turning to big data sources, such as mobile network operator (MNO) data, to supplement existing data. The paper focuses on utilizing synthetic MNO daily data from SIM cards issued to Ukrainian refugees by a Polish MNO. It proposes AMUSE model: mobility model for data deduplication and a simple estimator for estimating the present refugee population based on aggregated signalling data over time and areas. Further research shall be focused on including data variability over time into modelling.

## References

Ahas, R., Aasa, A., Yuan, Y., Raubal, M., Smoreda, Z., Liu, Y., Ziemlicki, C., Tiru, M. and Zook, M., (2015). Everyday space–time geographies: using mobile phone based sensor data to monitor urban activity in Harbin, Paris, and Tallinn. *International Journal of Geographical Information Science*, 29(11), pp. 2017–2039.

Alexander, L., Jiang, S., Murga, M. and González, M. C., (2015). Origin–destination trips by purpose and time of day inferred from mobile phone data. *Transportation Research Part C: Emerging Technologies*, 58, pp. 240–250.

Auerbach, F., (1913). Das gesetz der be volkerungskonzentration, Petermanns Geographische Mitteilungen, 59.

Baccelli, F., Błaszczyszyn, B., (2006). Tessellation in Wireless Communication Networks: Voronoi and Beyond it. Lorenz Center, Leiden University, 6-10 March 2006.

Badr, H., Du, H., Marshall, M., Dong, E., Squire, M. and Gardner, L., (2020). Association between mobility patterns and covid-19 transmission in the USA: a mathematical modelling study. The Lancet Infectious Diseases, 20(11).

Cleveland, R. B., Cleveland, W. S., McRae and J. E., Terpenning, I., (1990). STL: A Seasonal-Trend Decomposition Procedure Based on Loess. *Journal of Official Statistics*, 6, pp. 3–73.

Deville, P., Linard, C., Martin, S., Gilbert, M., Stevens, F. R., Gaughan, A. E., Blondel, V. D. and Tatem, A. J., (2014). Dynamic population mapping using mobile phone data. *Proceedings of the National Academy of Sciences*, 111(45), pp. 15888–15893.

Diao, M., Zhu, Y., Joseph Ferreira, J. and Ratti, C., (2016). Inferring individual daily activities from mobile phone traces: A Boston example. *Environment and Planning B: Planning and Design*, 43(5), pp. 920–940.

GUS, (2022). Health of refugees from Ukraine in Poland 2022, Household survey and behavioural insights research.

Lai, S., Erbach-Schoenberg, E., Pezzulo, C., Ruktanonchai, N., Sorichetta, A., Steele, J., Li, T., Dooley, C. and Tatem, A., (2019). Exploring the use of mobile phone data for national migration statistics. *Palgrave Communications 5*, 34.

Lanzieri, G., (2013). Population definitions at the 2010 censuses round in the countries of the UNECE region, in: 15th Meeting of the UNECE Group of Experts on Population and Housing Censuses, Geneva, Switzerland.

Lanzieri, G., (2019). Towards a single population concept for international purposes: definitions and statistical architecture, in: 16th Meeting of the Task Force on the Future EU Censuses of Population and Housing, Luxembourg.

Osborn, D., Chui, A., Smith, J. and Birchenhall, C., (1988). Seasonality and the order of integration for consumption. *Oxford Bulletin of Economics and Statistics*, 50(4), pp. 361–377.

Ricciato, F., Coluccia, A., (2021). On the estimation of spatial density from mobile network operator data. arXiv:2009.05410v3 [eess.SP].

Ricciato, F., Lanzieri and G., Wirthmann, A., (2020). Towards a methodological framework for estimating present population density from mobile network operator data. Pervasive and Mobile Computing, 68.

Ricciato, F., Widhalm, P., Craglia and M., Pantisano, F., (2016). Beyond the "single-operator, CDR-only" paradigm: An interoperable framework for mobile phone network data analyses and population density estimation. Pervasive and Mobile Computing.

Särndal, C-E., Swensson and B., Wretman, J., (1992). *Model Assisted Survey Sampling*, New York: Springer.

Saidani, Y., Bohnensteffen and S., Hadam, S., (2022). Quality Of Mobile Network Data – Project Experience And Use Cases In Official Statistics.

Salgado, D., Sanguiao, L., Onacea, B., et al. (2021). An end-to-end statistical process with mobile network data for official statistics. EPJ Data Sci. 10, 20(2021).

Shams, M., Kausar, N., Agarwal, P.and Oros, G. I., (2022). Efficient iterative scheme for solving non-linear equations with engineering applications. *Applied Mathematics in Science and Engineering*, 30:1, pp. 708–735, doi: 10.1080/27690911.2022.2130914.

Shepp, L. Vardi, Y., (1982). Maximum likelihood reconstruction for emission tomography. IEEE Transactions on Medical Imaging.

Tennekes, M., Gootzen, Y., (2021). A Bayesian approach to location estimation of mobile devices from mobile network operator data. *Journal of Spatial Information Science*.

UNHCR, (2023). Displacement Patterns, Protection Risks and Needs of Refugees from Ukraine, Regional Protection Analysis #3, Trends analysis: Moldova, Poland, Romania, and Slovakia, November 2023.

UNHCR, (2024). Ukraine Situation: Regional Refugee Response Plan, January-December 2024.

Urząd Komunikacji Elektronicznej, (2022). Raport o stanie rynku telekomunikacyjnego w Polsce w 2022 r.

Van der Laan, J., de Jongey, E., (2019). Maximum likelihood reconstruction of population densities from mobile signalling data. In NetMob'19.

Wesołowski, A. Eagle, N., Noor, A. M., Snow, R. W. and Buckee, C. O., (2013). The impact of biases in mobile phone ownership on estimates of human mobility. *Journal of the Royal Society*, Interface, Vol. 10(81).

Zagheni, E., Ingmar, W., (2012). You are where you E-mail: Using E-mail Data to Estimate International Migration Rates, WebSci 2012.

Zhu, Z., Klein, A. B., Li, G.and Pang, S., (2023). Fixed-point iterative linear inverse solver with extended precision. Sci Rep, 13(1):5198. https://doi.org/10.1038/s41598-023-32338-5.

Zipf, G. K., (1949). Human behavior and the principle of least effort: Cambridge, Massachusetts: Addison-Wesley.