Statistics Norway
Statistisk sentralbyrå

Li-Chun Zhang

Documents

Developing methods for determining the number of unauthorized foreigners in Norway

Statistics Norway/Division for Statistical Methods and Standards

Summary

This is a summary of the report on the first part of the joint research and development project funded by the Norwegian Directorate of Immigration (UDI, reference number 06/6594), developing methods for determining the numbers of unauthorized foreigners in Norway. The report is prepared by Senior Researcher Li-Chun Zhang.

A method is developed based on the data available. The expected total irregular residents population with non-EU origins is estimated to be 18196 by 1.1.2006. This constituted 0.39% of the official residents populations in Norway in 2005. The estimated lower and upper bounds of a 95% confidence interval are 10460 and 31917, respectively. Of the estimated total irregular residents, 12325 were previous asylum seekers, and the rest, 5871, were persons that had never applied for asylum. The estimated lower and upper bounds of a 95% confidence interval for the irregular residents excluding previous asylum seekers are 3352 and 10385, respectively.

Comparisons with relevant empirical results available from other EU countries suggest that the aforementioned estimates are plausible indicators of the size of the target population. It is, however, important to be aware of two basic preconditions for the results.

Firstly, there is the problem of data, which are few and difficult to extract. The details will be described in Section 3. The estimation approach is necessitated by the data that have been made available to us. Due to limited time and resources, it has not been possible to study in depth other potential data sources such as the many databases at the Police. Also, it has not been feasible to go more thoroughly through the data that may be available at UDI. A main difficulty is that the various types of potential data are not maintained for the purpose of this project, which makes test data extraction extremely costly and time-consuming. A more systematic and coordinated survey of the data sources should have a high priority in future methodological development.

Secondly, the estimation method makes use of certain model assumptions, explained in Section 5. It is important to realize that, while assumptions of various kinds are unavoidable in this case due to the nature of the problem, it is impossible to verify all the underlying assumptions beyond any empirical doubt. We therefore emphasize that one should not be overconfident in the reported estimates. Rather one should treat them as useful pieces of information that can help us towards a better overall understanding of the phenomenon of interest.

Below is a summary of the data and the model assumptions. For each country we have

m: number of unauthorized foreigners registered at UDI between May of 2005 and 2006.
n: number of foreign citizens who faced criminal charges during the calendar year 2005.
N: number of foreign-born persons of the age 18 and over, registered in the Central Population Register by 1.1.2006.

The data are prepared in aggregated table form such that no person can be identified. The aim is to estimate M, the number of irregular residents from each country, in order to arrive at the overall total by 1.1.2006. The model assumptions can briefly be described as follows:

(1) There is a relationship between M and N of the type: M is increasing with N, where the rate of increase decreases as N becomes larger. The actual rate of increase is unknown and estimated from the data.

(2) There is a relationship between m/M and n/N of the type: m/M is increasing with n/N, where the rate of increase decreases as n/N becomes larger. The actual rate of increase is independent of the rate of increase under (1) and is unknown and estimated from the data.

(3) In addition to the underlying structural assumptions in (1) and (2) we assume random variation from country to country.

The derived model for m with explanatory variables (n, N) fits well to the data, but can not verify beyond any doubt the assumed relationship between M and N. Statistical associations assumed in the model do not express any cause-effect relationships.

Jan Bjørnstad
Head of Research
Division for Statistical Methods and Standards
Statistics Norway

Contents

1 Introduction

This is the report on the first part of the joint research and development project funded by the Norwegian Directorate of Immigration (UDI, reference number 06/6594), developing methods for determining the numbers of unauthorized foreigners in Norway.

A method is developed based on the data available. The expected total irregular residents population with non-EU origins is estimated to be 18196 by 1.1.2006. This constituted 0.39% of the total residents population in Norway in 2005. The estimated lower and upper bounds of a 95% confidence interval are 10460 and 31917, respectively. Of the estimated total irregular residents, 12325 were previous asylum seekers, and the rest, 5871, were persons that had never applied for asylum. The estimated lower and upper bounds of a 95% confidence interval for the irregular residents excluding previous asylum seekers are 3352 and 10385, respectively.

Comparisons with relevant empirical results available from other EU countries suggest that the aforementioned estimates are plausible indicators of the size of the target population. It is, however, important to be aware of two basic preconditions for the results.

Firstly, there is the problem of data, which are few and difficult to extract. The details will be described in Section 3. The estimation approach is necessitated by the data that have been made available to us. Due to limited time and resources, it has not been possible to study in depth other potential data sources such as the many databases at the Police. Also, it has not been feasible to go more thoroughly through the data that may be available at UDI. A main difficulty is that the various types of potential data are not maintained for the purpose of this project, which makes test data extraction extremely costly and time-consuming. A more systematic and coordinated survey of the data sources should have a high priority in future methodological development.

Secondly, the estimation method makes use of certain model assumptions, explained in Section 5. It is important to realize that, while assumptions of various kinds are unavoidable in this case due to the nature of the problem, it is impossible to verify all the underlying assumptions beyond any empirical doubt. We therefore emphasize that one should not be overconfident in the reported estimates. Rather one should treat them as useful pieces of information that can help us towards a better overall understanding of the phenomenon of interest.

The rest of the report is organized as follows. A brief summary of the data sources that have been utilized internationally for unauthorized immigrants populations is presented in Section 2. Also provided are some relevant empirical results from other western countries. In Section 3 we describe the potential data sources in Norway, and the data that have been used in this project. The target parameter of estimation is defined and discussed in Section 4. We then outline in Section 5 a random effects mixed modeling estimation approach, and describe in Section 6 how it has been applied to yield the estimates of interest. Some topics for future developments are discussed in Section 7. Appendix A provides the technical details of the estimation method. For methodological comparisons, we provide in Appendix B and C critical reviews of the two existing scientific sample-based estimation methods, namely the truncated Poisson regression based on repeated captures data and the single stage link-tracing sampling.

2 A brief overview of international experiences

Below is a summary of data sources that have been reviewed by Pinkerton, McLaughlan, and Salt (2004) and Jandl (2004), with an emphasis on the European situation.

- Population data
  - **Census**: Unauthorized foreign-born persons can be counted in a traditional census with door-to-door visits. Subjected to a post-census adjustment for under- and/or over-counting, this provides an estimated population size at the moment of census.
  - **Register of foreign-born persons**: The immigration authority can be expected to maintain a register, with varying quality, of all foreign-born persons with permanent or temporary residence permit, as well as asylum seekers and refugees.
  - **Central population register (CPR)**: The CPR should ideally provide good coverage of long-term regular residents. Information on birth and death events of unauthorized residents can be expected, to a varying extent. A notable exception is Spain, where irregular immigrants have a strong incentive to register themselves. They are then eligible for social benefits such as free health care, while the data are not used for removing unauthorized residents from the country.
  - **Regularization program**: Regularization programs for undocumented migrants have been carried out in several European countries. Jandl (2004) identifies three weaknesses in the data. Firstly, not all illegal residents can or will take advantage of the regularization program. Secondly, persons who are granted a time-limited permit frequently fall back into the illegal status. Thirdly, the regularization program may generate temporary strong in-flows from neighboring countries.
  - Other **registers with partial coverage** such as a register of children at school.
- Sample data
  - **General large-scale survey**: There are several such surveys that provide at least partial coverage of the target population, including the Permanent Demographic Survey in France and the Labor Force Survey in a number of countries.
  - **Targeted survey**: There are two types of ultimate sampling units.
    - **Unauthorized immigrants** can be sampled/traced starting from sites (or institutions) where they are expected to be present with a high probability.
    - **Expert witnesses**: Knowledgable persons, such as officials or employers, may be asked to guess the target population size or, more likely, the proportion of irregular immigrants among their clients or branch of business. This type of data are subjective and the potential bias is ultimately not estimable.
  - **Apprehensions data**: It is often possible to separate between two sources:

* **Border apprehensions** of persons who attempt to enter the country illegally.
* **Common law enforcement** with apprehensions of illegal residents and workers.

Table 1: Estimated total numbers of unauthorized immigrants populations in some western countries and their approximate proportions to respective regular residents populations

Country	Total (in 1000)	Prop. (in %)	Source/Time of Reference/Brief Comments
USA	11500	3.8	Passel (2007), cf. References
EU	4500	≥ 1	Papademetriou, D.G. (2005). The "Regularization" Option in Managing Illegal Migration More Effectively: A Comparative Perspective. *Migration Policy Institute.*
Spain	700 - 800	1.7 - 1.9	Based on regularization programme in 2005.
	614	1.5	No. applications for regularization programme in 2001.
Italy	800	1.4	Workpermit.com (2006). Italy offers citizenship to illegal migrants after 5 years. *Estimate by the three largest Labor Unions CGIL, CISL and UILs.*
	700	1.2	Based on regularization programmes (2002-2003).
Netherlands	129	0.8	A 95% confidence interval: 74320 to 183912. van der Heijden *et al.* (2006). Een Scahtting van het aantal in Nederland verblijvende illegale vreemdelingen in 2005. *(In Dutch)*
UK	430	0.7	Lowest estimate 310000 and highest estimate 570000. Woodbridge, J. (2005). Sizing the unauthorized (illegal) migrant population in the United Kingdom in 2001. *Home Office Online Report 29/05.*
Austria	38.5	0.48	Apprehensions in 2004, incl. smuggled persons, and illegally entering and/or residing persons.
Sweden	31	0.35	Based on temporary program for previously rejected asylum seekers in 2005 - 2006. "Arbetet med den tilfalliga utlanningslagstiftningen 2005 - 2006". *(In Swedish)*

Table 1 contains references for some empirical results on the size of various unauthorized immigrants populations in the western countries. The following observations may be worth noting. To start with, the presence of unauthorized immigrants is clearly the highest in Mediterranean countries such as Spain and Italy. There is a gradual decrease as one moves towards central and northern Europe. Thus, a country's geographic distance towards the South, which is the main origin of unauthorized migration, is an important determining factor in the size of unauthorized immigrants population in that country. It follows that the numbers or, perhaps even more directly, the proportions of unauthorized immigrants in the neighboring Scandinavian countries can be expected to provide the best indications on the corresponding figures in Norway.

However, one needs to be very careful about the target population a number refers to. For instance, the Austrian figure consists only of the illegal immigrants that have actually been apprehended and, therefore, most likely it represents only a lower limit of the true target population size. Similarly, the Swedish figure consists only of previous asylum seekers. It does not cover many other groups of irregular immigrants, such as illegal border entries or illegal workers following legal entries. Also, it is not certain that all the applicants for the temporary program were residing in the country during the reference period. More on the definition in Section 4.

The above examples also clearly show the need for methodological developments and statistical modeling in order to handle the problem. It is unrealistic to expect the data sources to have such a coverage, and the associated qualities to be of such a standard, that the target number can be produced based on simple tabulations and calculations, as is perhaps the case with the population of regular residents.

3 Data sources in Norway

Below is a summary of relevant data sources in Norway that we are aware of.

- UDI maintains a so-called DUF-register. Individuals who have cases handled at UDI are assigned a DUF identification number.
- Statistics Norway
  - Census data: The last traditional census in Norway was in 1980. The census 1990 was a combination of census in the smallest municipalities and large samples in the larger ones. The last census in 2001 was post-/internet-based and directed only at dwellings. The population statistics were produced from statistical registers.
  - Statistics Norway maintains a statistical copy of the CPR. A registered person with residence permit for 6 months or more is assigned a person identification number. A person with less than 6 months residence permit is assigned a so-called D-number.
  - All regular surveys of persons and households at Statistics Norway, including the Labor Force Survey, use the CPR as the population frame. No systematic effort is made to survey the unauthorized immigrants should they be present in the selected household.
- The police has a number of databases where one may expect to find instances of unauthorized immigrants. However, the data are not organized in a way that allows test data extraction without considerable amount of time and resources. It was not possible to study the Police data in this project, apart from those that have made their way into the databases of UDI or Statistics Norway.
- Other governmental authorities or public institutions may have registers or databases that contain cases of unauthorized immigrants. These include e.g. the register of school children, the patients records at hospitals, *etc.* A major shortcoming of such data sources is that they provide only a partial coverage of the target population. No attempt was made in this project to extract data from them.
- Various humanitarian organizations may have contacts as well as records of unauthorized immigrants. There is a confidentiality issue of whether the relevant data can be made available for statistical purposes. No attempt was made to obtain such data.

Next, we summarize the data coverage of various types of foreigners.

- **Legal residents** The CPR provides a good coverage of legal foreign residents. Since UDI grants all the residence permits, it should also be possible to trace these persons in the DUF-register at UDI. However, the CPR is presumably more updated regarding events of birth, death and, probably to a less extend, emigration.

- **Quasi-legal residents** These are persons who are authorized to stay in the country yet do not have a regular residence permit. Typical examples are asylum seekers. The quasi-legal residents have DUF-numbers in the DUF-register.
- **Short-term visitors with DUF-number** These are tourists or business travelers who need entry visa. Such visitors are not counted as residents in standard practice.
- **Foreigners with free entry** These are legal residents in the EU member states covered by the Schengen Agreement, or other countries that have special free-entry agreement with Norway. Short-term visitors among them will not enter the DUF-register, but neither are they counted as residents.
- **Unauthorized immigrants** in the DUF-register due to expulsion requests are the only observed cases of unauthorized immigrants available to this project.
- Apprehensions of **legal foreign residents**
  - In theory an expulsion request should be filed by the police for any *sentenced* foreigners regardless of the residence status, by which apprehended legal foreign residents may enter the DUF-register. In practice, there appears to be a great under-count of such legal foreign residents in the DUF-register. A reason for this may be the fact that an expulsion request in the case of a legal immigrant is rarely granted except for heavy crimes such as murder or drug trafficking. This may have caused the police to drop the expulsion requests following minor crimes.
  - To a varying extent, foreigners who have been charged or sentenced for criminal offences can be traced in the statistical data that are provided to Statistics Norway by the police. Affirmation of irregular residence status is however difficult for persons who are not registered in the CPR.

Three main conclusions emerge from the above:

1. All available data at the moment are located either at UDI or Statistics Norway.
2. Potentially useful data may be found at the police, but only at extra costs.
3. Only sample-based estimation methods are currently possible in Norway, unless major policy changes lead to the collection of relevant population data.

The following data sets have been used for the calculation of the aforementioned estimates. The data are prepared in aggregated table form, such that the individuals behind the various counts can not be identified.

- The numbers of foreign-born persons by country of origin, who were of the age 18 or over, and were registered in the CPR on the 1st of January, 2006. Producer: Statistics Norway. (Remark: Ideally we would like to have more time to find out whether there is more suitable reference population that can be used; see discussions in Subsection 5.3.)

- The numbers of foreigners by country of citizenship, who faced criminal charges during the calendar year 2005. Producer: Statistics Norway. (Remark: These are the most reliable data of apprehended foreigners available to us. Ideally, we would like to be able to separate the irregular residents from the legal ones, which is not possible at the moment for a number of technical reasons. This does not invalidate the modeling approach as we shall explain in Section 5. But it is important to watch out for possible systematic changes in the data source, if the estimation procedure is to be applied repeatedly in future.)

- The numbers of foreigners by country of citizenship, who did not have a valid permit for staying in the country. Based on expulsion requests that have been handled at UDI during the 1st of May, 2005 and the 30th of April, 2006. The numbers are further divided into those who had applied for asylum and those who had not. Producer: UDI. (Remark: At the moment it is difficult to identify who has made an expulsion request in the DUF-register. The registration of such relevant information can be improved. Similarly, routines and information concerning the status of the regular residents may be strengthened.)

The motivations for extracting these data sets and the ways by which they have been utilized in the estimation will be explained in Sections 4 - 6.

4 Definition of target parameter

The population of unauthorized immigrants is both hard to access and hard to define.

To start with, a foreign-born person may be classified according to three characteristics:

- Entry status: legal or illegal
- Residence status: legal, quasi-legal, temporary or illegal
- Working status: legal, illegal or no-work

The exact definition of the categories varies from country to country, as well as from time to time, due to the dynamism and intricacies of the immigration laws.

In this project we focus on the residence status. Our target of estimation is the size of the irregular residents population in Norway, which is the term we will use from now on. The entry and working status are subordinate dimensions. Thus, an irregular resident may or may not have entered the country illegally, and he or she may or may not be working illegally. Due to the focus on the residence status, it was decided that citizens from the following countries will be excluded from the target population:

> Andorra, Austria, Belgium, Bulgaria, The Czech Republic, Cyprus, Denmark, Estonia, Finland, France, The Faroe Islands, Germany, Greenland, Greece, Hungary, Iceland, Ireland, Italy, Latvia, Liechtenstein, Lithuania, Luxembourg, Malta, Netherland, Poland, Portugal, Romania, Slovakia, Slovenia, Spain, Switzerland, Sweden, UK

Most of these are EU member states. The others are excluded because they are situated inside the EU zone and have a special diplomatic relationship with Norway.

Now, at any given moment, an irregular resident person must be someone who is present in the country without a legal basis. There are, however, several difficulties, of both practical and theoretical nature, that prevent us from adopting such a naturalistic (or physical) definition.

First of all, while it is in principle possible to implement the definition in a census-like mode of data collection with door-to-door visits, the associated cost is prohibitive. As reviewed above, observations of irregular residents in Norway, either at the Police or UDI, require accumulation of data over time. It is often impossible to verify the presence on a fixed reference time point of interest, such as the 1st of January, 2006. There is also a problem with person identification, which is necessary in order to distinguish between potential multiple observations and/or stays of the same person. Moreover, the legality status of a person may change, sometimes several times, over the period of data collection, and one rarely succeeds in tracing the whole history.

Next, it seems that one needs to distinguish between residence and presence. Thus, technically speaking there can be many kinds of transitory or trivial illegal presence of foreigners, but not all of them are, or should be, counted as irregular residence. For example, many applications for asylum are filed from within the country, which means that technically there is a post-entry period when the applicant has no legal ground for being present in the country. To commit oneself to

the naturalistic definition above would have required going through all such possibilities, without realistic hopes of resolving the problems given the quality of the relevant data.

Moreover, our target population is hardly static. Thus, for example, in a hypothetical situation with round-the-clock complete surveillance, we would know exactly how many irregular residents there are, respectively, on the first and second of January, 2006. The two numbers will almost surely be different. But such spurious variations are hardly of interest from our point of view in a project like this one. It can therefore be argued that a theoretical stable size is preferable in order to avoid this kind of accidental random variation.

Finally, the so-called residual method (e.g. Passel, 2007) is undoubtedly the most well studied and documented estimation procedure available at the moment. The essential relationship is summarized in the following equation:

$$U = A - L$$

where U stands for the total of unauthorized immigrants, and A stands for that of all immigrants, and L stands for the total of legal immigrants. Various adjustments of the terms on the right-hand side are necessary due to death events, emigrations, census under-counts, *etc.*. In reality only theoretical or expected values can be calculated, but not the actual figures for the particular reference period of concern. It follows that the method yields an estimate U that is of an implied, theoretical and stable nature, unlike the naturalistic definition above.

In light of these general remarks, we shall now define our target parameter as follows. Let M be the size of the irregular residents population at the time point of interest. Let N be the size of a known *reference* population at the same time point. For the estimates reported in this report, N is the number of foreign-born persons of age 18 or over, who were registered in the CPR on the 1st of January, 2006. Consider M as a random variable and N a known covariate. Denote by $f(M|N)$ the conditional probabilistic distribution of M given N. The target parameter is the *theoretical* size of irregular residents, which is defined as the conditional expectation of M given N with respect to $f(M|N)$, denoted by

$$\xi = E(M|N) \tag{1}$$

The following observations may be worth noting. First, the theoretical size is defined to be the conditional expectation of a random variable. This enables us formally to get rid of the spurious variation as long as the reference population size is held fixed. Next, the introduction of N serves two purposes: (a) it can be used as an explanatory variable of the irregular size M, and (b) it provides an interpretation of the irregular size M in analogy to N. For example, part of the idea of a theoretical size is to have a stable measure of the target population size, where the time-dependent variation in M is linked to that of N. Provided the day-to-day variation in N around the 1st of January in 2006 can be considered immaterial, the theoretical size ξ is valid not just for the 1st of January in 2006, but for a period around it. Moreover, since the chosen N is not subjected to great seasonal variations, neither is the theoretical ξ. In comparison, based on the

naturalistic definition, M is supposed to vary greatly from one time of the year to another, being perhaps the highest in the summer months, which is another kind of spurious variation that is not of importance to us.

5 A random effects mixed modeling approach

5.1 Some remarks on the data

In addition to M and N, let m be the number of irregular residents that were observed between May, 2005 and April, 2006, and let n be the number of foreign citizens who faced criminal charges in the calendar year 2005. Both have been explained before at the end of Section 3. The idea now is to develop a statistical model, which makes it possible to estimate M based on the observed N, n and m.

The accumulation of observations implies that m and n must be collected over a period of time. A calendar year seems a natural choice if the target parameter (1) is to be estimated on a yearly base. Technically speaking, however, the two types of cases do not have to be collected from the same period, or even over two periods of the same length, as long as it is possible to find a statistical model that works for the data available. Although we would have liked to examine alternative choices of the reference periods, such an option did not exist for us.

In concept, n should be a *reference* count that can plausibly be related to both m and N. With m being the number of apprehended irregular foreigners and N the number of legal foreign residents, a natural choice of n would be the number of apprehended legal foreign residents. However, as explained before, it is not possible for us to decide in all the cases whether a foreigner charged with a crime is a legal resident or not. In fact, one can be quite certain that some of these persons are irregular residents. Yet we have no choice but to use the only data that are available. However, it is important to realize that this does not cause bias in the estimator of M. In contrast, bias would have been the case had m contained legal foreign residents. Moreover, whether a statistical model using the available n is acceptable or not is a matter that can be examined empirically by looking at the actual goodness-of-fit of the model.

It is clear from the above remarks that n/N and m/M are not literally the catch rates among the reference and target populations, respectively. Moreover, they must differ for a number of reasons such that a simplistic model like $E(m/M|n, N) = n/N$ surely can not hold. Below we shall develop a random effects mixed modeling approach that makes it possible to estimate M under more general conditions.

5.2 A hierarchical Poisson gamma model

For both the target and the reference populations, let $i = 1, ..., t$ be the index of the sub-population classified by the country of citizenship and origin, respectively. Assume that the observed number of irregular residents follows a Poisson distribution, with parameter λ_i, denoted by

$$m_i \sim \text{Poisson}(\lambda_i) \tag{2}$$

It is intuitively plausible that the parameter λ_i should depend on two other quantities: (a) the total number of irregular residents from country i, denoted by M_i, and (b) the probability of

being observed, i.e. the probability for an irregular resident to be in the DUF-register, denoted by p_i, i.e. $\lambda_i = M_i p_i$. In addition, let $u_i = M_i p_i / E(M_i p_i | n_i, N_i)$, where $E(M_i p_i | n_i, N_i)$ denotes the conditional expectation of $M_i p_i$ given n_i and N_i. The u_i is a random effect that accounts for heterogenous variation from one country to another. Together, we have

$$\lambda_i = \mu_i u_i \qquad \text{where} \qquad \mu_i = E(M_i p_i | n_i, N_i) = E(M_i | N_i) \cdot E(p_i | M_i, n_i, N_i) \tag{3}$$

We complete the model specification by assuming that

$$\xi_i = E(M_i | N_i) = N_i^{\alpha} \tag{4}$$

$$E(p_i | M_i, n_i, N_i) = E(p_i | n_i, N_i) = (\frac{n_i}{N_i})^{\beta} \tag{5}$$

$$u_i \sim \text{Gamma}(1, \phi) \tag{6}$$

where $\text{Gamma}(1, \phi)$ denotes the gamma distribution with expectation $E(u_i) = 1$ and variance $V(u_i) = 1/\phi$. Together, formulae (2) - (6) define a hierarchical Poisson gamma model.

We note that the model implies that $E(m_i/M_i | M_i, n_i, N_i) = (n_i/N_i)^{\beta}$. The term hierarchical refers to the fact that random variation exists on two different levels. Firstly, on the population level, the Poisson parameter λ_i depends on u_i that is a random gamma-variable from one country to another. In contrast, μ_i contains the fixed effects α and β that are constant across all the countries. Secondly, given u_i (and, thus, λ_i), the observation m_i is subject to a random sampling error following the corresponding Poisson distribution.

Notice that we have modeled the expected Poisson parameter $\mu_i = E(\lambda_i)$ as a product of equations (4) and (5). Given the data available, this is necessary in order to incorporate the irregular size M_i into the model, as we are not interested in a model that only explains the observed counts m_i, for $i = 1, ..., t$. The parameters α and β are identifiable, and the combined model for μ_i, i.e. the product of equations (4) and (5), can be tested empirically for its goodness-of-fit to the observed data. Having done that, we may use the estimate of α to derive an estimate of $\xi_i = E(M_i | N_i)$. However, in a fundamental sense, the plausibility of equation (4), or (5), can not be empirically established on its own. To put it in another way, it is conceivable that one may be able to come up with another model equation for $\mu_i = E(\lambda_i)$ that fits equally well to the observed counts m_i, for $i = 1, ..., t$. To choose between two alternative models in such a situation, one must rely *a priori* on assumptions that can not be verified by the data directly. This is the main reason that we have warned previously against overconfidence in the reported estimates. Meanwhile, it is equally important to point out that we have not found any decisive evidence that speaks *against* the assumptions (4) and (5). Theoretical motivations for the proposed model are given below in Subsection 5.3, whereas the empirical results will be documented in Section 6.

5.3 On model assumptions

First of all, it seems natural that M_i should depend on the reference population size N_i. The lack of common social-health benefits and regular job opportunities means that an irregular resident needs a network of contact in order to survive. Moreover, the contact circle would most likely contain regular residents who have a more solid social-economic fundament. It is hard to imagine that completely closed community of irregular residents can be the norm of existence in a society like the Norwegian one. This explains the choice of the reference population, namely, foreign-born persons with age 18 or over. It seems reasonable to believe that this group should contain most of the *direct* contact with the irregular residents from the same country of origin. People with foreign roots who were born in Norway can only have contact with the country of origin through their parents or elder relatives, whereas the contact a foreign-born person has would not cease to exist although the person by now may hold a Norwegian citizenship. Finally, it seems plausible that only adults among them can act as dependable resources for the irregular residents.

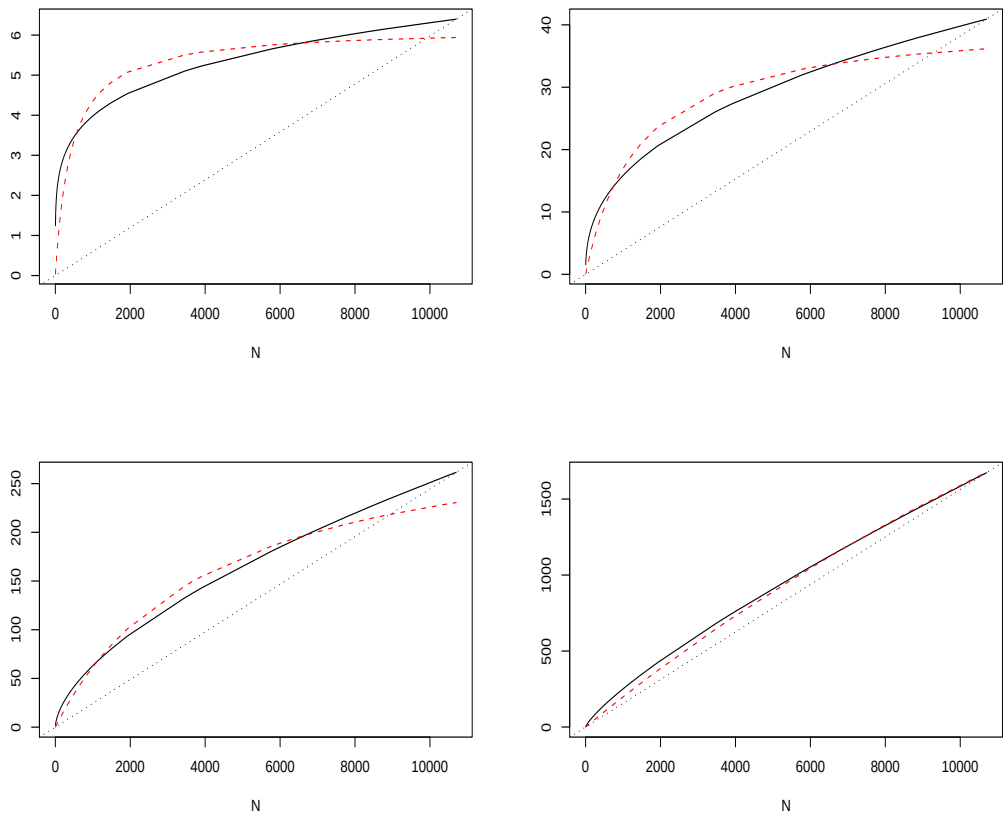

Figure 1: Illustrations of power curve $\xi = N^\alpha$ (solid) and Beveton-Holt curve $R = \nu S/(1+\kappa S)$ for $S = N$ (dashed). Dotted lines for linear development. Top-left: $\alpha = 0.2$ and $(\nu, \kappa) = (0.015, 0.002388)$. Top-right: $\alpha = 0.4$ and $(\nu, \kappa) = (0.029, 0.000707)$. Bottom-left: $\alpha = 0.6$ and $(\nu, \kappa) = (0.076, 0.000238)$. Bottom-right: $\alpha = 0.8$ and $(\nu, \kappa) = (0.203, 0.000028)$.

In equation (4) we assume that the expected irregular size ξ_i is given by N_i^α. We expect α to have a value between 0 and 1, in which case ξ_i increases more quickly for small N_i and, then, gradually flattens out as N_i becomes large. This is known as the density dependence property that e.g. are commonly found in ecological models for wild-life animal populations. For instance, the Beverton-Holt model (Beverton and Holt, 1957) is given by

$$R = \frac{\nu S}{1 + \kappa S}$$

where S: current size of reproduction basis

R: offsprings in the next generation

κ : density dependence parameter

ν/κ : carrying capacity, i.e. a reproduction "ceiling" as S tends to infinity

The Beverton-Holt curve has a similar density dependent development as the power curve in equation (4), as S and N tend to infinity. The density dependence parameter κ regulates the rate of convergence towards the carrying capacity ν/κ, where a large κ yields quicker convergence. A constant linear relationship would be the case without the density dependence parameter, i.e. $R = \nu S$ provide $\kappa = 0$. It is possible to conceive the reference population of regular residents as the "existence basis", and the irregular residents as a particular kind of "offsprings" that depend on the reference population. Four illustrations of the power curve N^α and the Beverton-Holt curve are given in Figure 1. Also shown is the constant linear development for comparison. It is seen that both curves exhibit the density dependence property. Moreover, similar developments over a particular range of the independent variable can be expressed using either model curve with suitably chosen parameters. The main difference is that the Beverton-Holt model assumes an asymptotic limit ν/κ for the dependent variable, whereas such a limit does not exist under equation (4) as ξ_i goes to infinity together with N_i. It is difficult to speculate about the exact nature of such asymptotic behaviors, but an absolute population "ceiling" of irregular residents does not appear necessary in our case.

Next, because by equation (5) we are modeling a rate, n_i/N_i is a natural choice of explanatory variable. Notice that we may rewrite the equation (5) on the log scale as

$$\log E(p_i|n_i, N_i) = \beta \log(n_i/N_i)$$

Both rates n_i/N_i and m_i/M_i are similar to proportions of binary outcomes. The logistic transformation would be the standard choice for such quantities, where e.g.

$$\text{logit}(p_i) = \log\{p_i/(1-p_i)\} = \log(p_i) - \log(1-p_i)$$

The log transformation is preferable because it enables us to combine the two model equations

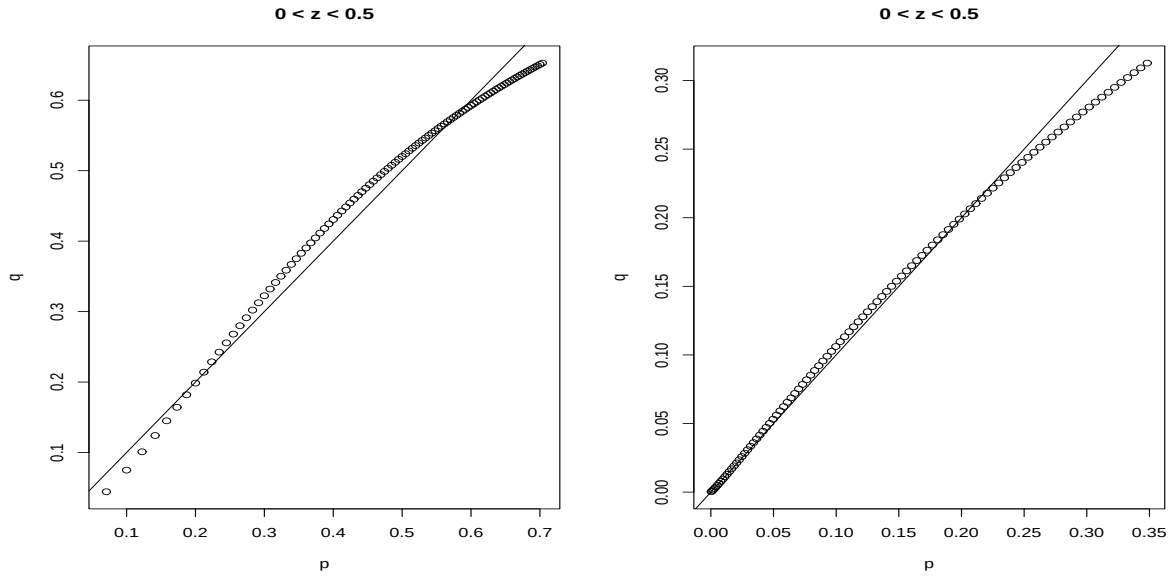

Figure 2: Illustrations of log and logit link functions. Left: $\log(p) = 0.5\log(z)$ and $\text{logit}(q) = 1.197 + 0.806\log(z)$. Right: $\log(p) = 1.5\log(z)$ and $\text{logit}(q) = 0.345 + 1.611\log(z)$.

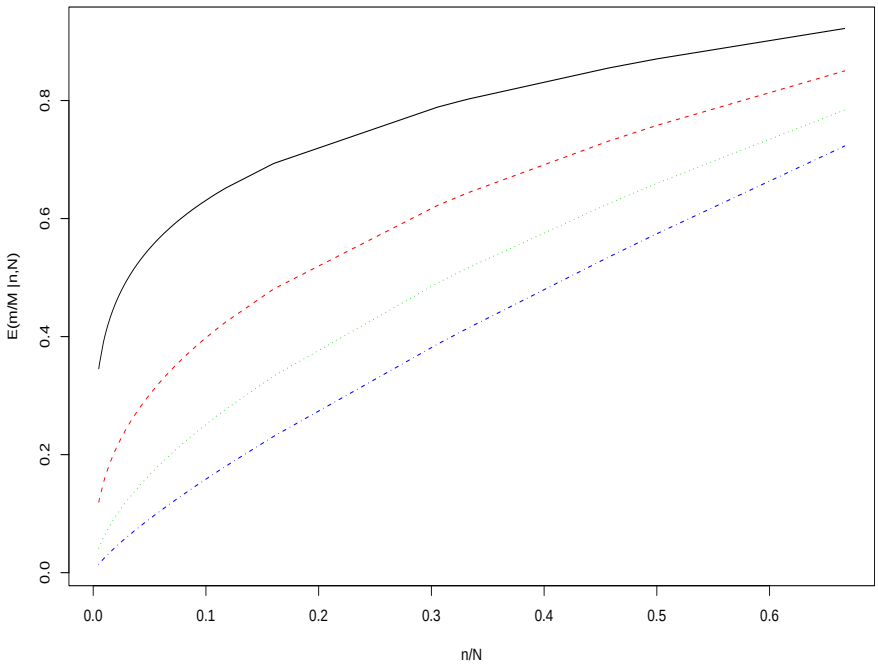

Figure 3: Illustrations of $E(m_i/M_i|n_i, N_i) = (n_i/N_i)^\beta$ for $\beta = 0.2$ (solid), $\beta = 0.4$ (dashed), $\beta = 0.6$ (dotted) and $\beta = 0.8$ (dotted and dashed), and actual n_i/N_i in the data, arranged in the increasing order.

(4) and (5) directly on the same scale to yield

$$\mu_i = N_i^{\alpha}(\frac{n_i}{N_i})^{\beta} \tag{7}$$

Moreover, we expect the rates to be small. For any $\pi \approx 0$, we have $\pi/(1-\pi) \approx \pi$ and $\text{logit}(\pi) \approx \log(\pi)$. Also, the two link functions can be quite similar even though it is not directly the case that $\log(\pi) \approx \text{logit}(\pi)$. Two illustrations are given in Figure 2. In each case we plot two probabilities against each other, denoted by q and p. Here we have a covariate z. On the y-axis a specific logistic model is assumed for q, while on the x-axis a specific log model is assumed for p. Figure 2 shows that for a given log link function, there exists a corresponding logistic link function which yields very similar probabilities as the covariate z varies for $0 < z < 0.5$. Finally, Figure 3 illustrates the model equation (5), i.e. $E(p_i|n_i, N_i) = E(m_i/M_i|n_i, N_i) = (n_i/N_i)^{\beta}$, for $\beta = 0.2, 0.4, 0.6, 0.8$. The values of n_i/N_i are the actual ones in the data available, arranged in the increasing order.

It seems plausible that the underlying assumptions for μ_i should be confined to the country of origin, and observations by country of origin make the parameters identifiable, as well as creating useful degrees of freedom in the data to allow us to estimate the precision of the resulting estimator. The introduction of the random effects by formula (6) is motivated by several important considerations. The equations (4) and (5) give us the expectation of the Poisson parameter λ_i, i.e. an overall relationship that is valid throughout the target population. It is, however, realistic to assume that there is variation from one country to another that makes the actual λ_i deviate from its expected value. Such heterogenous variation is accounted for through the random effects u_i. The assumption of gamma distribution for the random effect u_i is common in combination with the Poisson distribution of the observed count m_i. Together they form a conjugate family of distributions, in the sense that the conditional distribution of u_i given m_i is again a gamma distribution. This is very convenient for computation. The random effect u_i is restricted to have unity mean, similar to assuming zero mean for the residuals in a linear regression model. Otherwise it would imply mis-specification of the model equation for μ_i. It is possible to check on the unity-mean assumption empirically, as we will do in Section 6.

6 Estimation results

6.1 Summary of results

In this Section we apply the Poisson gamma model to derive the reported estimates. The data to be used have been defined at the end of Section 3. Recall also that the target population does not include citizens from the 33 countries listed in Section 4. The main results are summarized in Table 2 below. The relevant details and remarks are given in the sequel.

Table 2: Estimates of theoretical size $\xi = E(M|N)$ and parameters (α, β, ϕ). Lower and upper bounds of 95% confidence interval for ξ. Standard errors of parameter estimates in parentheses.

Target Irregular Residents Population	m	n	N	$\hat{\xi}$	Lower	Upper
All	1959	5747	152496	18196	10460	31917
Without Previous Asylum Seekers	941	5747	152496	5871	3352	10385
Previous Asylum Seekers	1018	5747	152496	7631	3286	18118
	-	-	-	12325†	-	-

	Parameter Estimate (Standard Error)		
Target Irregular Residents Population	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\phi}$
All	0.742 (0.035)	0.692 (0.073)	3.617 (0.834)
Without Previous Asylum Seekers	0.603 (0.037)	0.599 (0.078)	5.156 (1.376)
Previous Asylum Seekers	0.641 (0.054)	0.624 (0.115)	1.218 (0.253)

The results are divided in three parts: (a) the total number of all irregular residents on January the 1st, 2006, (b) the number of irregular residents who have never applied for asylum, (c) the number of irregular residents who have been asylum seekers at some point in time. Of course,

$$\xi(a) = \xi(b) + \xi(c)$$

The estimates that are derived directly for each target population (a) - (c) do not satisfy this condition. As explained in Section 6.5, the two models with the most explanatory power for the given covariates are for the populations (a) and (b), yielding 12325 (= 18196 − 5871, and marked by † in Table 2) as the estimated number of irregular residents among previous asylum seekers.

Notice that the population (b) is further divided into 3 subgroups at UDI: (b.1) persons who have been granted a visa on false grounds, (b.2) persons who have overstayed beyond the expiring date of a visa or living permit - including those who initially do not need a visa to enter the country, and (b.3) persons who entered the country without a required visa and without being registered. We do not provide estimates for these subgroups.

6.2 On the estimation of total number of irregular residents

In the available data we had 175 countries with $N_i > 0$. Among these 92 countries satisfy the following conditions: (i) $m_i > 0$, (ii) $n_i > 0$, and (iii) $n_i/N_i < 1$. We group the remaining 83 countries together and create a pseudo-country, denoted by *Rest*. In this way the countries that initially do not satisfy the conditions (i) - (iii) can now be treated just like any other 'observed' data point, and it is possible to check empirically whether this leads to a worsened fit of the model. In Figure 4 we have plotted m_i against n_i for these 93 data points. The pseudo-country Rest (marked by "+") appears to fit reasonably well into the pattern of the data. There is a clear outlier (marked by "X"), where the number of observed irregular residents is way beyond what can be expected based on the other countries. The outlier is left out, and the estimation is based on the remaining 91 countries, and the pseudo-country Rest.

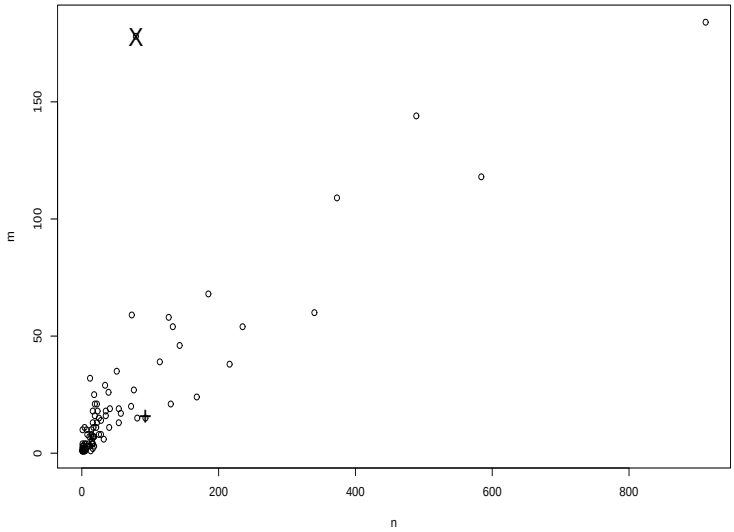

Figure 4: Scatter plot of m_i against n_i. Outlier (X) and Rest (+).

Equation (7) implies on the log scale that $\log(\mu_i/N_i) = (\alpha - 1)\log N_i + \beta \log(n_i/N_i)$. Hence, a model for $\log(m_i/N_i)$ is of the form

$$\log(m_i/N_i) = (\alpha - 1)\log N_i + \beta \log(n_i/N_i) + \epsilon_i \tag{8}$$

This provides easy means for exploring the model equation (7). In Figure 5 we have plotted $\log(m_i/N_i)$ against $\log(N_i)$ and $\log(n_i/N_i)$, respectively. In both cases there is a clear marginal linear relationship (marked by the dotted lines), as can be expected from the model assumptions underlying (7). Again, the pseudo-country Rest (marked by "+") appears to fit reasonably well into the pattern of the data, while the outlier looks misplaced.

Diagnostics plots on fitting the Hierarchical Poisson gamma model to the data depicted in

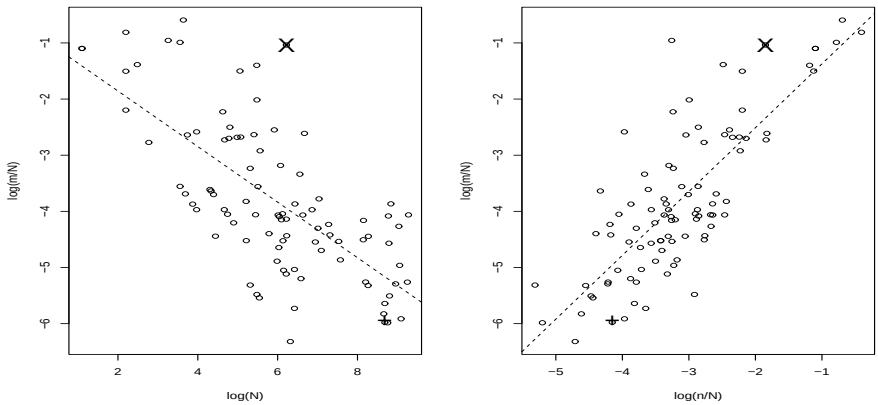

Figure 5: Exploration of model for μ_i. Rest (+). Outlier (X). Marginal linear relationship (dotted).

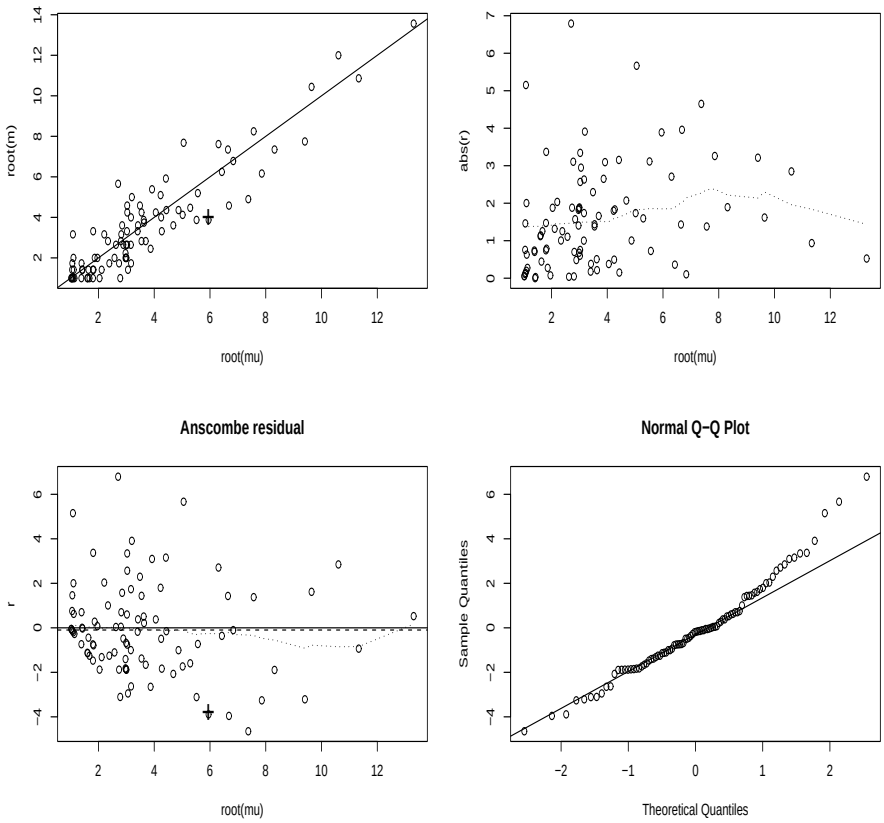

Figure 6: Model diagnostics for estimation of total irregular residents. Rest (+).

Figure 4 and 5 are given in Figure 6. In the top-left corner is the scatter plot of $\sqrt{m_i}$ against $\sqrt{\hat{\mu}_i}$, i.e. its estimated mean value. The square root transformation is used to make the points more evenly spaced over the axes. The solid line marks the equality line should these two be identical. There is no indication of bias in the model specification of μ_i.

In the bottom-left corner, the Anscombe residuals are plotted against $\sqrt{\hat{\mu}_i}$. The Anscombe residual for the Poisson distribution is given by

$$r_i = \frac{3m_i^{2/3} - 3\hat{\mu}_i^{2/3}}{2\hat{\mu}_i^{1/6}}$$

see e.g. McCullagh and Nelder (1989). The idea is to transform a non-normal random variable to a scale where the normal approximation is the best. The overall mean of the Anscombe residuals is marked by the dashed line, which is rather close to zero (marked by the solid line). In addition, the dotted line shows the running means of the residuals. Suitably specified predictor μ_i should yield a flat curve that is close to zero, which appears to be case here.

In the top-right corner the absolute Anscombe residual $|r_i|$ is plotted against $\sqrt{\hat{\mu}_i}$, together with the running means (marked by the dotted line). This provides a check of the variance assumption. Large curvature of the running means would indicate deviations from the underlying model assumption. There is no strong indication for this in the current plot.

Finally, the Q-Q normal plot in the bottom-right corner shows that the distribution of the Anscombe residuals is quite normal on the lower side of the mean, but somewhat long-tailed on the upper side. But the extent of the deviations from normality is not extreme enough to cast serious doubts on the results. In summary, the Poisson gamma model yields a reasonable fit to the data, and the pseudo-country Rest (marked by "+" in the left-hand side plots) does not constitute an abnormal data point.

The estimated number of non-EU irregular residents for all countries is 18196 (Table 2), and a 95% confidence interval is given as $(10460, 31917)$. The parameter estimates of (α, β, ϕ) and their standard errors are given in the second block there. All the parameters are highly significant. We notice the followings.

- The estimated total is 18196. This is 0.39% of the total population in Norway in 2005, which seems plausible compared to the proportion of 0.35% for irregular residents who were previous asylum seekers in Sweden (Table 1). Still, for reasons explained before, we would like to caution against overconfidence in this particular point estimate. On the other hand, the importance of the confidence interval should not be overlooked.

- The data, the estimate and a 95% confidence interval for the pseudo-country Rest are given in Table 3. Notice that $\hat{\xi}_{Rest}/N_{rest} = 627/5913 = 0.106$ is slightly below the overall ratio $\hat{\xi}/N = 18196/152496 = 0.119$, which seems plausible. In addition, ξ is estimated to be 17802 without the pseudo-country Rest in the data, which differs by 1.3% from the estimate by subtraction, i.e. $18196 - 627 = 17569$. The estimation appears to be robust towards the

way these rest countries are handled.

Table 3: Data, estimate and 95% confidence interval for Rest and Outlier.

	m	n	N	$\hat{\xi}$	Lower	Upper
Rest	15	93	5913	627	343	1146
Outlier	178	79	502	101	65	155

- A similar summary is given for the outlier country in Table 3. Being an outlier to the model, $\hat{\xi}_{outlier}$ would not be a good prediction of the actual $M_{outlier}$. Indeed, it can be seen that $\hat{\xi}_{outlier} = 101$ is actually smaller than the observed $m_{outlier} = 178$. It should be mentioned that it is well known that the data in this country have background in certain organized illegal activities. It is reasonable to believe that this abnormal situation will not continue over time. Consequently, it can be argued that the temporary upsurge in the number of irregular residents from this outlier country should not affect the estimation of our target parameter, especially with regard to its interpretation as a theoretical, stable measure that changes smoothly over time.

6.3 On the estimation of irregular residents excluding previous asylum seekers

Among 175 countries with $N_i > 0$ there are 79 that satisfy all the following conditions: (i) $m_i > 0$, (ii) $n_i > 0$, and (iii) $n_i/N_i < 1$. As before we group the remaining countries together and create a pseudo-country, denoted by *Rest*. In Figure 7 we have plotted m_i against n_i for these 80 data points. Again, there is a clear outlier (marked by "X"), which is the same country as above. The outlier is left out, and the estimation is based on the remaining 79 data points.

In Figure 8 we have plotted $\log(m_i/N_i)$ against $\log(N_i)$ and $\log(n_i/N_i)$, respectively. In both cases there is a clear marginal linear relationship (marked by the dotted lines). The pseudo-country Rest (marked by "+") appears to be somewhat outlying. It is thus important to keep track of its impact on the results. The outlier again looks misplaced. Diagnostics plots on fitting the Poisson gamma model are given in Figure 9. Overall the goodness-of-fit appears acceptable. In particular, the pseudo-country Rest (marked by "+" in the left-hand side plots) does not appear overly abnormal.

The estimated number of non-EU irregular residents excluding previous asylum seekers is 5871 (Table 2). A 95% confidence interval is given as $(3352, 10385)$. All the parameter estimates are highly significant. We notice the followings.

- The estimated total without previous asylum seekers constitute about 1/3 of the estimated total of all irregular residents. No estimate of this particular sub-population has previously been available in Norway. There is very little empirical ground for comparison. However, the Schengen agreement has made it relatively easy for irregular immigrants to move across a large territory stretching from the Mediterranean shores to the Scandinavian coasts, in

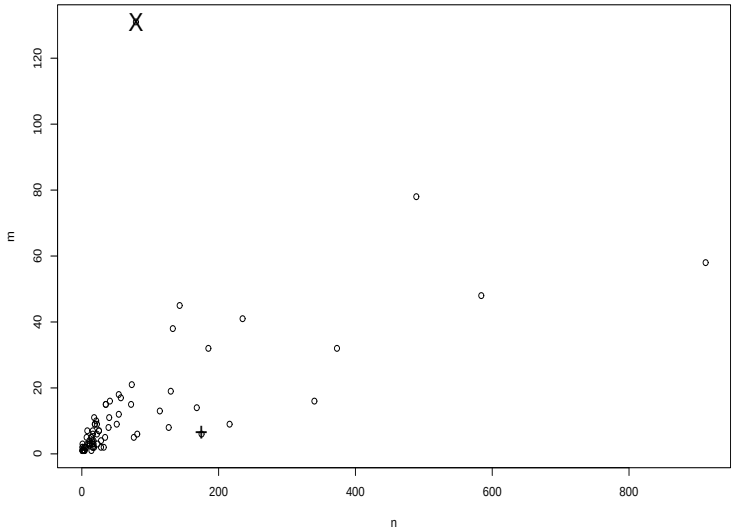

Figure 7: Scatter plot of m_i (without previous asylum seekers) against n_i. Outlier (X) and Rest (+).

the light of which the reported magnitude does not seem unlikely. Again, we would like to emphasize the importance of the confidence interval.

- A summary is given for the pseudo-country Rest in Table 4. Notice that $\hat{\xi}_{Rest}/N_{rest} = 225/7971 = 0.028$ lies below the overall ratio $\hat{\xi}/N = 5871/152496 = 0.038$. In addition, ξ is estimated to be 5868 without the pseudo-country Rest in the data, which differs by 3.9% from the estimate by subtraction, i.e. $5871 - 225 = 5646$. While the impact is somewhat larger than in the case of all irregular residents in Section 6.2, the estimation remains fairly robust towards the way the rest countries are handled.

Table 4: Data, estimate and 95% confidence interval for Rest and Outlier.

	m	n	N	$\hat{\xi}$	Lower	Upper
Rest	6	175	7971	225	118	428
Outlier	131	79	502	42	27	66

- Also given in Table 4 is a summary of the outlier country. Again, $\hat{\xi}_{outlier} = 42$ is way below the observed $m_{outlier} = 131$. The same considerations apply as in Section 6.2.

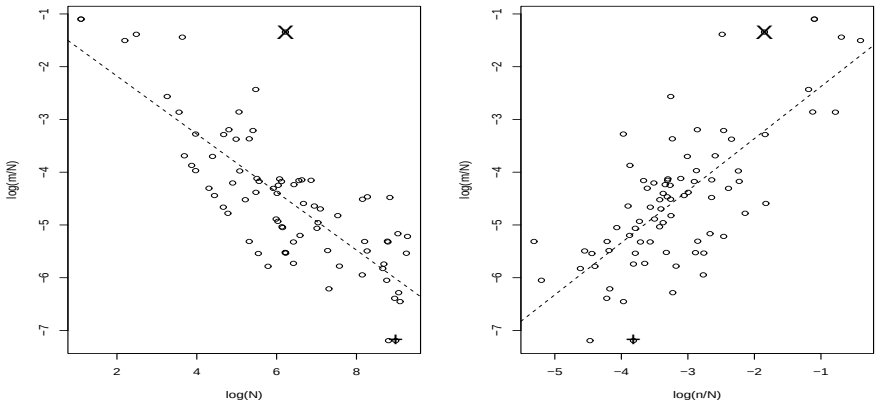

Figure 8: Exploration of model for μ_i without previous asylum seekers. Rest (+). Outlier (X). Marginal linear relationship (dotted).

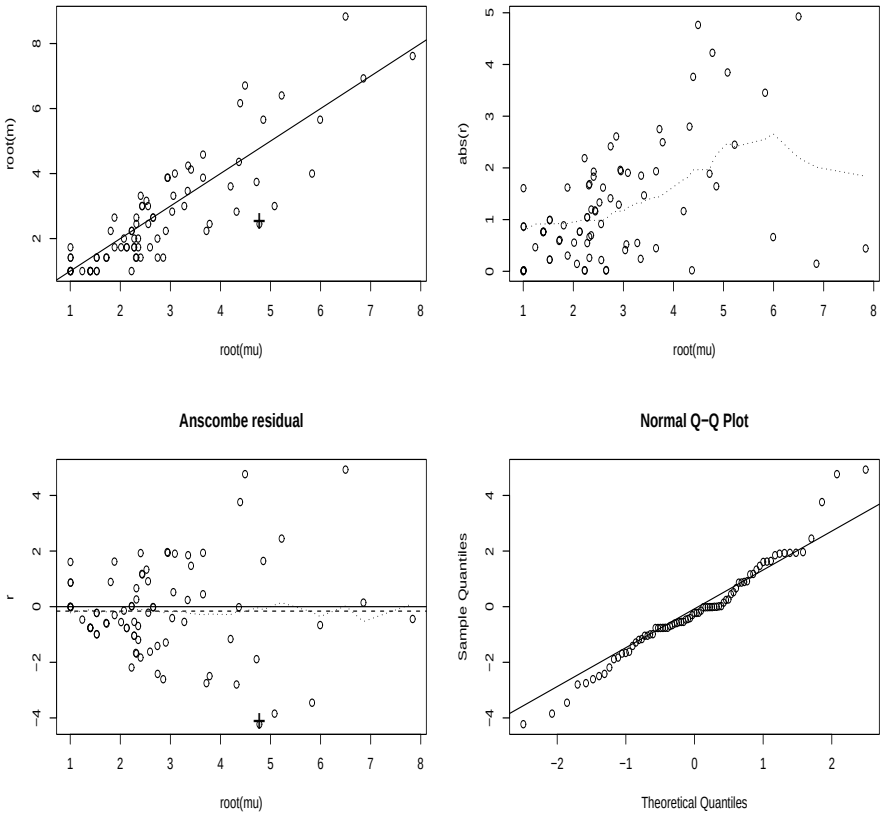

Figure 9: Model diagnostics for estimation of total without previous asylum seekers. Rest (+).

6.4 On the estimation of irregular residents among previous asylum seekers

Among 175 countries with $N_i > 0$ there are 74 that satisfy all the following conditions: (i) $m_i > 0$, (ii) $n_i > 0$, and (iii) $n_i/N_i < 1$. As before we group the remaining countries together and create a pseudo-country, denoted by *Rest*, giving altogether 75 data points.

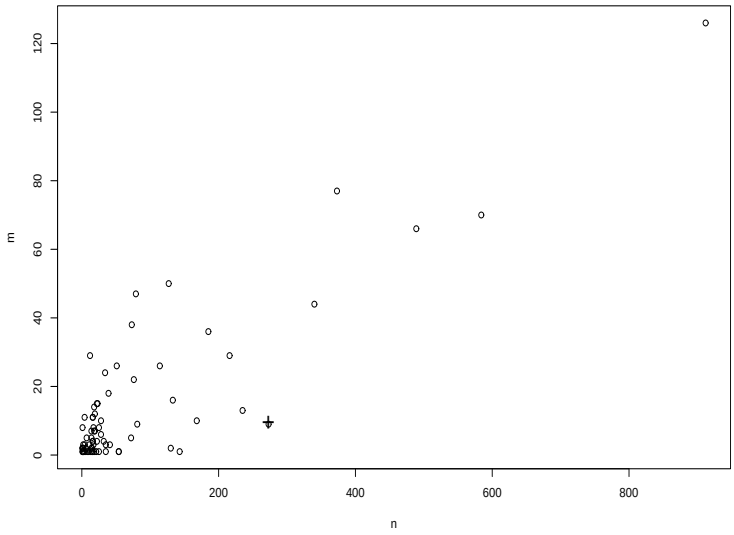

Figure 10: Scatter plot of m_i (previous asylum seekers) against n_i. Rest (+).

The various plots are given in Figure 10 - 12. They were no clear outliers at first sight. The estimated number of irregular residents among previous asylum seekers is 7631, and a 95% confidence interval is given by $(3286, 18118)$; see Table 2. By closer look at the left-hand side plots of Figure 12 we see that there are several countries with fairly large n_i but very small m_i. In addition, the pseudo-country Rest appears somewhat outlying. We have thus carried out the estimation again, where the pseudo-country Rest was removed from the observations, together with the other two countries that had the largest negative Anscombe residuals in the bottom-left plot of Figure 12. The resulting estimate of the total and the parameter estimates are given in Table 5. The estimate of the total is raised by about 15%. Finally, there is the indirect estimate by subtraction, i.e. $\hat{\xi}(c) = \hat{\xi}(a) - \hat{\xi}(b) = 18196 - 5871 = 12325$, also given in Table 2.

Table 5: Alternative estimate for total irregular residents among previous asylum seekers and related parameter estimates. Lower and upper bounds of 95% confidence interval. Standard errors of parameter estimates in parentheses.

Previous Asylum Seekers			Parameter Estimate (Standard Error)		
$\hat{\xi}$	Lower	Upper	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\phi}$
8810	3877	20431	0.659 (0.053)	0.644 (0.111)	1.432 (0.319)

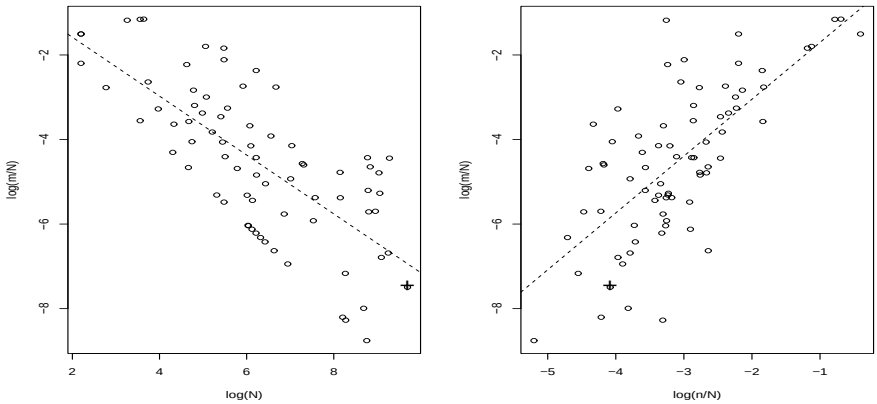

Figure 11: Exploration of model for μ_i (previous asylum seekers). Rest (+). Marginal linear relationship (dotted).

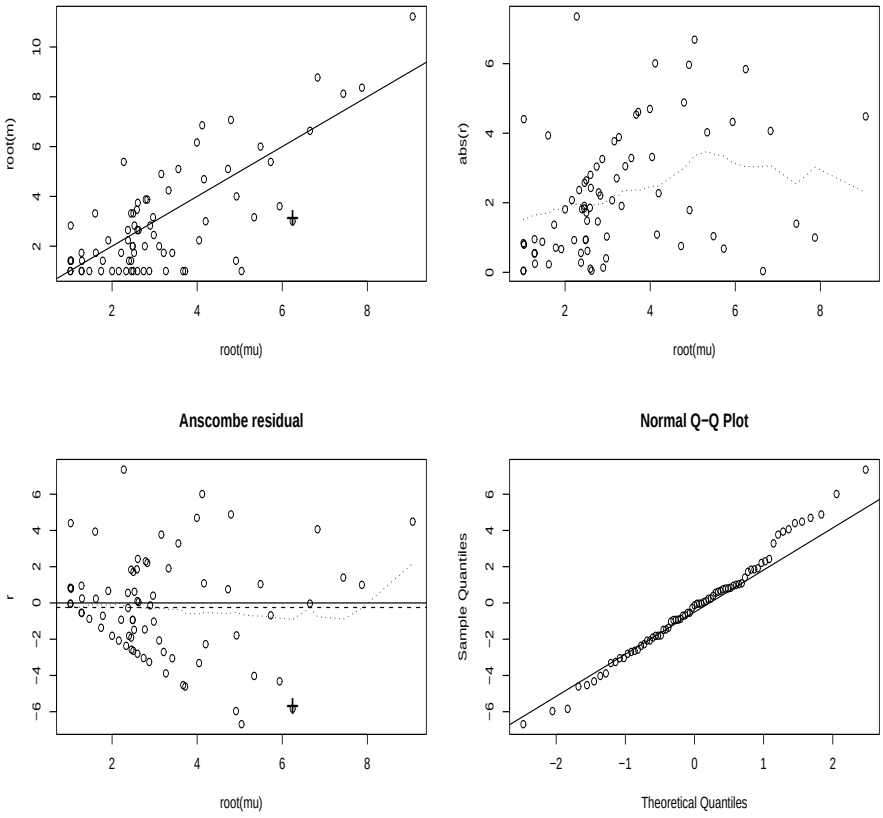

Figure 12: Model diagnostics for estimation of total among previous asylum seekers. Rest (+).

6.5 Discussion

Clearly, the main difference in the estimates of $\xi(c)$ is between the indirect and the direct estimates. This is due to the non-linear model equation (4). Generally speaking, for any given α_1 and α_2, there does not exist another constant α such that $N^{\alpha} = N^{\alpha_1} + N^{\alpha_2}$ for all N. Thus, the model equation (4) is not additive in the sense that we can use it to model $\xi(a)$, $\xi(b)$ and $\xi(c)$ all at once using the *same* reference population size. It follows that the inconsistency between the indirect and direct estimates can not be solved based on the data that are currently available to us, and we have to choose, at most, two of the three direct estimates and derive the third one indirectly based on the two chosen estimates.

Table 6: Three alternative sets of estimated totals.

Scenario (I)			Scenario (II)			Scenario (III)		
$\hat{\xi}(a)$	$\hat{\xi}(b)$	$\hat{\xi}(c)$	$\hat{\xi}(a)$	$\hat{\xi}(b)$	$\hat{\xi}(c)$	$\hat{\xi}(a)$	$\hat{\xi}(b)$	$\hat{\xi}(c)$
Direct	Direct	Indirect	Direct	Indirect	Direct	Indirect	Direct	Direct
18196	5871	12325	18196	10565	7631	13502	5871	7631

The three alternative choices are shown in Table 6. The following considerations seem relevant.

- One way to check the explanatory power of the three directly estimated Poisson gamma models is to look at the estimated variance of the random effect, i.e. $\widehat{V}(u_i) = 1/\hat{\phi}$. The smaller the estimate $\hat{\phi}$, the larger is the variance of the random effects, which means that less of the variation in the data is explained by the given covariates. From Table (2), $\hat{\phi}(a) = 3.617$, $\hat{\phi}(b) = 5.156$ and $\hat{\phi}(c) = 1.218$. The model for the irregular residents among previous asylum seekers is quite clearly the least powerful one.

- Scenario (II) appears unlikely since $\hat{\xi}(b) > \hat{\xi}(c)$. It is conventional wisdom that previous asylum seekers constitute the majority among the irregular residents. Take for instance Sweden in Table 1. If $\xi(b) > \xi(c)$, then the proportion of all irregular residents would have been raised to at least the same level as Netherlands and UK. This seems implausible given the more tightly regulated society and the geographic location of Sweden.

- Some measures of the numbers of previous asylum seekers are given in Tables 7 and 8. It is seen in Table 7 that in the five years between 2003 and 2007 there could have been up to 16676 previous asylum seekers who were present in the country after the deadline by which they should have left. Of course, not all of them were present on the 1st of January, 2006. Also, there could be many irregular residents among the 18475 persons in the category Others. An indication of this can be seen in Table 8. Among the 5465 persons who had applied for asylum in the year 2005, there were 1994 who were rejected and obliged to leave the country. The registered expulsion rate is only $325/1994 = 0.163$. The registered expulsion rate is even lower in the category Others, which comprises of persons whose applications were either not considered or were rejected without further consideration. In other words, an

absolute majority of persons who were not granted asylum were potential irregular residents, and there is no indication that the number of irregular residents who had been asylum seekers could not be as high as 12325 in Scenario (I).

Table 7: Categories of previous asylum seekers by year of final decision of applications.

Year	2003	2004	2005	2006	2007	Total
Asylum Granted	3135	3965	2989	2151	3962	16202
Observed Over-stayer*	5640	4070	2777	2438	1751	16676
Others	5003	5169	2759	1862	3682	18475
Total	13778	13204	8525	6451	9395	51353

Table 8: Categories of asylum seekers who had applied in year 2005

		Rejected and Obliged to Leave		
Total	Granted	Expulsion Registered	Expulsion Unregistered	Others
5465	1790	325	1669	1681

∗ *An over-stayer has had the asylum application rejected. The deadline date for leaving the country is set at four weeks after the final decision if unregistered. After the deadline date, an over-stayer is observed if she/he (i) had a registered date for leaving the country, or (ii) had been located at a detention center, or (iii) had a registered private address. The registration data in the cases of (i) - (iii) have a high quality. The uncertainty is mainly associated with the deadline data in cases it is calculated.*

In summary, we have chosen to report the results of Scenario (I). In doing so we have based ourselves on the two models that have the most explanatory power, yielding 18196 as the estimated total number of irregular residents and 5871 for irregular residents excluding previous asylum seekers. There is no decisive empirical evidence against 12325 as the derived total of irregular residents among previous asylum seekers. Indeed, this derived indirect estimate falls comfortably within the directly estimated confidence interval $(3286, 18118)$, in Table 2.

7 Some topics for further development

We have developed a general random effects mixed modeling approach for sizing the irregular residents population in Norway. Three important features are worth noting:

- The target parameter (1) is defined as the theoretical size of irregular residents population, instead of a naturalistic definition of the actual size.
- The introduction of a known reference population of the size N and a known reference count n as explanatory variables for the observed count of irregular residents.
- The use of random effects that allow for heterogenous variation beyond what can be accounted for by the fixed covariates.

While there are strong theoretical motivations for all these choices, it is equally important to notice that the existing alternative sample-based estimation methods are simply not feasible based on the data that are currently available in Norway; see Appendix B and C.

Although the modeling approach has been sufficiently established, it may be possible to improve and refine the actual model adopted for the estimation. In particular, this concerns the choice of N and n. For instance, is it possible to use some suitably defined total of asylum seekers as the reference size N for irregular residents who had previously been asylum seekers? What should then be the choice of n? Would it be better to apply different models in each of the sub-populations?

We are only beginning to utilize the potential information in the data that are scattered around in different systems and in various forms. For instance, it may be possible to find better reference counts n in the databases at the police. There are most likely other relevant information in the DUF-register that are useful either directly or indirectly. It may be possible to strengthen the registration of such relevant information, and improve the organization of the databases, in order to make it easier to extract the information for statistical purposes. At least one should develop standardized routines for data extraction. Are there other valuable data sources apart from the ones that have been identified for this project? A systematic and fruitful survey of the various potentially useful data is crucial for further methodological developments.

It may be possible to adapt the current modeling approach to similar problems. For instance, to develop a model that can be used to estimate the number of illegal workers. Of course, this would require one to go through all the issues that we have dealt with in this report for the population of illegal workers.

A Estimation method

The target parameter and its estimator are given as, respectively,

$$\xi = \sum_{i=1}^{t} E(M_i|N_i) = \sum_i N_i^{\alpha} \qquad \text{and} \qquad \hat{\xi} = \sum_i N_i^{\hat{\alpha}}$$

where $\hat{\alpha}$ is the estimator of α. We shall use the maximum likelihood estimator (MLE). Denote by $L(\eta; \mathbf{m})$ the likelihood of $\eta = (\alpha, \beta, \phi)$ given m_i, for $i = 1, ..., t$. Under the Poisson gamma model (2) - (6), we have

$$f(m_i, u_i; \eta) = \frac{e^{-\mu_i u_i}(\mu_i u_i)^{m_i}}{m_i!} \cdot \frac{\phi^{\phi} u_i^{\phi-1} e^{-\phi u_i}}{\Gamma(\phi)} = \frac{\mu_i^{m_i}\phi^{\phi}}{m_i!\Gamma(\phi)} e^{-u_i(\mu_i+\phi)} u_i^{m_i+\phi-1}$$

where $\Gamma()$ is the gamma function. Thus,

$$\begin{aligned} f(m_i; \eta) &= \int_0^{\infty} f(m_i, u_i; \eta) d(u_i) \\ &= \frac{\mu_i^{m_i}\phi^{\phi}}{m_i!\Gamma(\phi)} \int_0^{\infty} e^{-(\sqrt{u_i})^2(\mu_i+\phi)} (\sqrt{u_i})^{2(m_i+\phi-1)} 2\sqrt{u_i}\, d(\sqrt{u_i}) \\ &= \frac{\mu_i^{m_i}\phi^{\phi}}{m_i!\Gamma(\phi)} (\mu_i + \phi)^{-(m_i+\phi)} \Gamma(m_i + \phi) \end{aligned}$$

based on the identity $\int_0^{\infty} e^{-\gamma z^2} z^k dz = \frac{1}{2}\gamma^{-\frac{k+1}{2}}\Gamma(\frac{k+1}{2})$, with $z = \sqrt{u_i}$ and $k = 2(m_i + \phi) - 1$. Notice that, conditional on m_i, u_i has the gamma distribution with mean $(m_i + \phi)/(\mu_i + \phi)$ and variance$(m_i + \phi)/(\mu_i + \phi)^2$.

The likelihood is given by

$$L(\eta; \mathbf{m}) = \prod_{i=1}^{t} f(m_i; \eta)$$

The log-likelihood is thus, disregarding constant terms, given by

$$l(\eta; \mathbf{m}) = \sum_{i=1}^{t} l_i(\eta)$$

where

$$\begin{aligned} l_i(\eta) &= m_i \log \mu_i - (m_i + \phi)\log(\mu_i + \phi) + \log\Gamma(m_i + \phi) + \phi \log \phi - \log\Gamma(\phi) \\ &\doteq m_i \log \mu_i - (m_i + \phi)\log(\mu_i + \phi) + \phi \log \phi \\ &\quad + (m_i + \phi - 0.5)\log(m_i + \phi) - (m_i + \phi) - (\phi - 0.5)\log(\phi) + \phi \\ &= m_i \log \mu_i - (m_i + \phi)\log(\mu_i + \phi) + (m_i + \phi - 0.5)\log(m_i + \phi) + 0.5 \log \phi \end{aligned}$$

by the Stirling approximation, $\log\Gamma(z) \doteq (z - 0.5)\log(z) + 0.5\log(2\pi) - z$.

The mean parameter μ_i is linear on the log scale, denoted by $\log \mu_i = x_i^T \gamma$ with generic vector of covariates x_i and parameters γ. Now that $l_i(\eta)$ depends on γ only through μ_i, we have

$$\frac{\partial l_i(\eta)}{\partial \gamma} = \frac{\partial l_i(\eta)}{\partial \mu_i} \frac{\partial \mu_i}{\partial \log \mu_i} \frac{\partial \log \mu_i}{\partial \gamma} = \frac{\partial l_i(\eta)}{\partial \mu_i} \mu_i x_i = \frac{m_i - \mu_i}{\mu_i + \phi} \phi x_i$$

where $\partial l_i(\eta)/\partial \mu_i = m_i/\mu_i - (m_i + \phi)/(\mu_i + \phi)$, and

$$\frac{\partial l_i(\eta)}{\partial \phi} = -\log(\mu_i + \phi) - \frac{m_i + \phi}{\mu_i + \phi} + \log(m_i + \phi) + \frac{m_i + \phi - 0.5}{m_i + \phi} + \frac{1}{2\phi}$$

Moreover,

$$\frac{\partial^2 l_i(\eta)}{\partial \gamma \partial \gamma^T} = \frac{\partial^2 l_i(\eta)}{\partial \mu_i^2} \mu_i x_i \frac{\partial \mu_i}{\partial \gamma^T} + \frac{\partial l_i(\eta)}{\partial \mu_i} x_i \frac{\partial \mu_i}{\partial \gamma^T} = -(\frac{m_i + \phi}{\mu_i + \phi} \mu_i \phi) x_i x_i^T$$

$$\frac{\partial^2 l_i(\eta)}{\partial \phi^2} = -\frac{2\mu_i + \phi - m_i}{(\mu_i + \phi)^2} + \frac{m_i + \phi + 0.5}{(m_i + \phi)^2} - \frac{1}{2\phi^2}$$

$$\frac{\partial^2 l_i(\eta)}{\partial \gamma \partial \phi} = (\partial(\frac{\partial l_i(\eta)}{\partial \mu_i})/\partial \phi) \mu_i x_i = -\frac{\mu_i - m_i}{(\mu_i + \phi)^2} \mu_i x_i = (\frac{\partial^2 l_i(\eta)}{\partial \phi \partial \gamma^T})^T$$

The MLE of η, denoted by $\hat{\eta}$, is given by the solution to the likelihood equations, i.e.

$$\frac{\partial l(\eta; \mathbf{m})}{\partial \eta} = \sum_{i=1}^{t} \frac{\partial l_i(\eta)}{\partial \eta} = 0$$

The MLE can be obtained using the Newton-Raphson method. As the starting values we use the ordinary least square fit of the heuristic log-ratio model (8). We use the estimated α and β as the starting values for the same parameters of the Poisson-gamma model, and the inverse of the estimated $V(\epsilon_i)$ as the starting value for ϕ. The asymptotic variance-covariance matrix of the MLE is given by the inverse of $-\partial^2 l/\partial \eta \partial \eta^T$, evaluated at $\eta = \hat{\eta}$. A confidence interval (CI) for ξ can be obtained directly by plugging in the CI for α, the latter of which can be based on the asymptotic normal approximation. That is, let (c_1, c_2) be a CI for α. The corresponding CI for ξ is then given as $(\sum_i N_i^{c_1},\ \sum_i N_i^{c_2})$.

B Repeated captures method

Recently, Van der Heijden, Cruyff, and Van Houwellingen (2003) and Van der Heijden, Bustami, Cruyff, Engbersen, and Van Houwellingen (2003) proposed a truncated Poisson regression approach to estimation of hidden populations, based on repeated apprehensions data in the police records. Below we provide a brief overview of the approach, and discuss some perceived obstacles that need to be resolved.

B.1 A brief overview

In the simple case the data can be arranged as a vector, denoted by N_1, N_2, N_3, ..., where N_k is the number of persons that have been apprehended for k offences. Let $N_{obs} = \sum_{k\geq 1} N_k$ be the number of persons who have been caught at least once. Denote by N_0 the hidden target population, i.e. persons who have committed offences but are never apprehended. The total target population size is given by $N = N_0 + N_{obs}$.

Assume that the number of apprehensions of a member of the target population, denoted by y_i for $i = 1, ..., N$, follows a Poisson distribution with parameter λ. We have

$$P(y_i = 0) = e^{-\lambda} \quad \text{and} \quad P(y_i | y_i > 0; \lambda) = \frac{P(y_i; \lambda)}{P(y_i > 0; \lambda)} = \frac{e^{-\lambda}\lambda^{y_i}}{y_i!(1 - e^{-\lambda})}$$

The parameter λ can be estimated based on the observed N_1, N_2, ..., denoted by $\hat{\lambda}$, under the truncated Poisson distribution. This yields then $\hat{N}_0 = N_{obs}e^{-\hat{\lambda}}/(1 - e^{-\hat{\lambda}})$ and $\hat{N} = \hat{N}_0 + N_{obs}$. Alternatively, an *estimated* Horvitz-Thompson estimator is given by

$$\hat{N}_{HT} = \frac{N_{obs}}{1 - e^{-\hat{\lambda}}} = \sum_{i=1}^{N_{obs}} \frac{1}{P(y_i > 0; \hat{\lambda})}$$

Van der Heijden, Bustami, Cruyff, Engbersen, and Van Houwellingen (2003) extended the simple setting above under a truncated Poisson regression model. For $i = 1, ..., N_{obs}$, put

$$\log(\lambda_i) = x_i^T \beta$$

where x_i is a column vector of covariates, and β contains the regression coefficients. Again, the parameters β can be estimated based on N_1, N_2, ..., denoted by $\hat{\beta}$, giving $\hat{\lambda}_i$ for each apprehended persons. The parameter λ_i remains unknown outside the sample, for $i = 1, ..., N_0$, unless x_i is known throughout the population. A Horvitz-Thompson type estimator is proposed, where

$$\hat{N}_{HT} = \sum_{i=1}^{N_{obs}} \frac{1}{P(y_i > 0 | x_i; \hat{\beta})} = \sum_{i=1}^{N_{obs}} \frac{1}{1 - \exp(-x_i^T \hat{\beta})}$$

B.2 On contagion

Two main assumptions have been discussed in the aforementioned papers. The first one is that the sample count of each member of the population follows a Poisson distribution. A perceived potential problem is known as contagion from the biostatistical literature: positive contagion is the case if previous apprehensions increase the probability of subsequent apprehensions; whereas negative contagion is the case if the probability decreases.

It is hard to reject contagion completely, because it would imply that apprehension has no effect at all on the behavior of the apprehended, nor the police. However, the interpretation that contagion necessarily leads to violation of the Poisson distribution needs some consideration. Obviously, this is true in cases of extreme contagion. For instance, apprehension may lead to exclusion from the population, as when an apprehended irregular immigrant is effectively expelled. In this way, extreme contagion is related to the issue of an open or closed population.

Next, if as a result of each apprehension the probability of being caught later is increased, then contagion may eventually lead to violation of the Poisson distribution. This can be explained by the genesis of the Poisson distribution as the limit of a binomial distribution with probability p and M trials, when M tends to infinity and p tends to zero in such a way that Mp tends to λ. A necessary condition here is that p should be very small, or close to zero.

In case the probability of being caught later decreases following previous apprehensions, or if the probability increases but remains very small, contagion does not necessarily leads to violation of the Poisson distribution. The probability of the binomial distribution does not have to be constant for the Poisson limit to hold. Thus, the total count of two independent binomial trials with parameters (M_1, p_1) and (M_2, p_2), respectively, can be approximated by the Poisson distribution with parameter $\lambda = \lambda_1 + \lambda_2 = M_1 p_1 + M_2 p_2$, provided each binomial distribution can be approximated by a Poisson distribution with parameter λ_1 or λ_2. Contagion implies merely that λ_2 depends on λ_1, but not dependence between the two binomial counts.

Suppose we split the full period of data collection into sub-periods between each apprehension, and in each sub-period a number of independent Bernoulli trials take place with a sub-period specific probability that can be approximated by a Poisson distribution with sub-period specific parameter. The number of apprehensions still follows a Poisson distribution for each person, but the parameter will be different for people who are caught different numbers of times. In this way contagion may lead to a particular kind of heterogeneity in the data, which is similar to informative sampling *with* replacement. Unless one is able to specify the dependence between the sub-period parameters correctly, estimation will be biased if contagion is ignored.

B.3 More on heterogeneity

The second assumption of the truncated Poisson regression approach concerns potential heterogeneity in the individual Poisson parameters. In the simple case, the Poisson distribution is identical throughout the whole population. Under the truncated Poisson regression model, the

distribution is identical for persons with the same covariates. The homogeneity assumption is violated if there are differences in the individual Poisson parameters that can not be explained by the observed covariates. As we have argued above, heterogeneity may be caused by contagion. To discuss the matter in general terms, we believe it is helpful to make explicit two concepts that give rise to the Poisson parameter.

Over the given period of data collection, let M be the *number of exposures*, i.e. when a member of the target population is susceptible to the force of law, and let p be the *hit rate*, i.e. the probability to catch an exposed member of the population. For example, in the biological context where animals are captured repeatedly, the number of exposures can be the number of times the catchers are out in the field, and the hit rate is the probability of an animal being caught on each of those occasions. The point is that outside the field working days the animals are not exposed, i.e. not susceptible to be captured.

Van der Heijden, Cruyff, and Van Houwellingen (2003) applied the truncated Poisson regression model to estimate two hidden populations, namely, drunken drivers and persons who illegally possess firearms. However, while a person violates the law all the time she/he is in possession of illegal firearms, no one is drunk and thus violates the law every time she/he drives. So how is the population of drunken drivers defined? Anyone who has ever driven while being drunk? Now, a driver is only exposed at those times when she/he actually drives while being drunk. Since the number of exposures varies in the population, the Poisson parameter must be heterogeneous in the population beyond, say, what can be explained by age, sex and region. Meanwhile, the motivations given for estimating the population size of drunken drivers are (i) insights into the threat this population may pose on society, and (ii) measure of the workload of the police. On both accounts it seems more reasonable to define the target population as cases of drunken driving rather than drunken drivers. Yet, in order to arrive at an estimate of total cases of drunken driving based on the number of people who have ever driven while being drunk, one needs the distribution of exposures in the population. In either case the estimation seems ill-conceived.

What about an *existential* population such as the illegal firearm-owners? It seems that there are different kinds of exposure and hit rate. A person may actually use illegal firearms in a criminal act, or a person may occasionally carry an illegal weapon around just for its own satisfaction, or a person may keep illegal firearms at home all the time. Clearly, the number of exposures and the associated hit rate are quite different in these situations. The result is again a heterogenous population of Poisson parameters, but for a reason different than contagion.

B.4 Two further remarks

An important motivation for introducing the truncated Poisson regression model is that it provides a means to account for the heterogenous Poisson parameters through the covariates available. The proposed Horvitz-Thompson type estimator is motivated as follows. Suppose a person with covariates x is apprehended, and the probability that a person with these covariates is apprehended is estimated to be $\hat{p}_x = P(y > 0|x; \hat{\beta})$. Then, for this person, one may estimate that there are

$\hat{p}_x^{-1} - 1$ other persons outside the sample that have the same x. For this argument to hold, however, there must be other people with the same x outside the sample. Thus, for example, it is wrong when Van der Heijden, Cruyff, and Van Houwellingen (2003) include in the regression model a covariate like "age of first offence", because if this actually means "age of first apprehension" as we have reasons to believe, then the variable exists only inside the sample. For the same reason, the truncated Poisson regression model should not include as covariates anything about a person's recorded criminal history. But it is allowed to distinguish between different types of exposure and hit rate, say, by the type of offences.

A more difficult problem arises when the target population is clearly not closed during the data collection period, such as that of the irregular immigrants. On the one hand, assumptions of a stable and closed population appear necessary. On the other hand, how can one reconciliate obviously different estimates based on samples collected in periods of *different* lengths, now that they all aim at the same number? In general, referring to the concepts of exposure and hit rate, people with different life durations in the population should have different Poisson parameters. This is yet another source of heterogeneity that needs to be taken into account.

C Single-stage link-tracing sampling

Link-tracing sampling (LTS), or snow-ball sampling, is a method that has been used for estimation of hidden and hard-to-access human populations such as drug users and homeless people. The idea is to enlarge the sample by asking the already-sampled persons to nominate other members of the target population. Successive waves are created if the nominees are subsequently asked to make nominations again. The sampling is terminated either according to some pre-specified stopping rules, or if new nominees cease to emerge. In particular, the single-stage LTS stops immediately after the persons in the initial sample have made their nominations.

In a wide range of situations the estimation is only possible using model-based methods. See Thompson and Frank (2000) for a review. Even then the requirement on data can be difficult to meet in practice. In particular, it may be difficult to identify, let alone to follow up, on the nominees, because the target population is hard to access. Here we shall consider only single-stage LTS. Each person in the initial sample is asked to make nominations of other members of the target population. The nominees are not to be identified, except for one or two most basic classification variables, whose details will be given as we proceed. Notice that the low risk of disclosure may encourage cooperation from the initial sample.

C.1 A graph model

Frank and Snijders (1994) studied a simple graph model. Consider a directed graph whose vertices represent the members of the target population, denoted by $U = \{1, ..., N\}$ with unknown N. Denote by $s = \{1, ..., n\}$ the initial sample of persons. Denote by an *ordered* pair (i, j), for $i, j \in U$, an arc from i to j, if the person j is nominated by the person i. Denote by A all the arcs, i.e. nominations, that come out of the sample s, i.e. $A = \{(i, j); i \in s\}$. The data of the single-stage LTS consist then of s and A. Moreover, we shall assume that the arcs of A are not identified, except for whether $j \in s$ or not. That is, we know the identity of each person in the initial sample, but for the nominees only whether they are in the initial sample or not. A variation of the situation is when the nominee j is known to belong to a certain group of the target population.

Frank and Snijders (1994) made two basic assumptions. (I) Assume that the initial sample s arises from Bernoulli sampling from U, with unknown selection probability p. (II) Conditional on s, assume that the arc set A arises from Bernoulli sampling from all $n \times (N-1)$ possible arcs in $\{(i, j); i \in s \text{ and } i \neq j\}$, with unknown selection probability ψ. In the case where the nominees are not completely identified, but can be classified according to whether $j \in s$ or not, let m_i be the number of nominations by i, and let r_i be the number of nominations by i that fall inside the initial sample. Let $m = \sum_{i \in s} m_i$ and $r = \sum_{i \in s} r_i$ be the respective sample totals.

We have, conditional on n, a moment estimator of N from the following two equations

$$r = n(n-1)\psi \quad \text{and} \quad m - r = n(N-n)\psi \qquad \Rightarrow \qquad \tilde{N} = n + (n-1)(m-r)/r$$

The estimator takes a prediction form, i.e. the observed number n plus the predicted number $(n-1)(m-r)/r$ outside the initial sample, where the latter can be motivated by simple expansion based on the proportional relationship $\{n(N-n)\}/\{n(n-1)\} = (m-r)/r$, which yields $(n-1)(m-r)/r$ as an estimate of $N-n$.

As a numerical example, Frank and Snijders (1994) used the data from a study of cocaine use in Rotterdam. The initial sample consisted of $n = 34$ persons. The total number of nominations were $m = 311$, of which $r = 15$ pointed into the initial sample. These yielded then $\tilde{N} = 685$.

The above graph model rests on two Bernoulli assumptions. Bernoulli sampling (I) of the initial sample may be questionable, if the actual sampling procedure "catches" some members of the target population more easily than the others. Félix-Medina and Thompson (2004) considered a generalization based on cluster sampling of *sites*, where the members of the target population have a high probability of being "caught". Bernoulli sampling (II) of the nominees presumes connectivity of the graph, i.e. it is possible for a person i to nominate *any* member of the target population other than him-/herself. This is hard to motivate unless the target population is narrowly confined, either geographically or by other means.

C.2 A variation of the graph model

In order to allow different persons to have varying *contact circle*, with possibly different numbers of potential nominees, we now postulate a variation of the graph model.

(i) Assume Bernoulli sampling of the initial sample, denoted by $n \sim Bin(N, p)$.
(ii) Assume independent Poisson distribution of nominations m_i, denoted by $m_i \sim Pois(\lambda)$.
(iii) Conditional on m_i, assume Binomial distribution of r_i, i.e. $r_i \sim Bin(m_i, (n-1)/(N-1))$.

The Poisson distribution (ii) of nominations does not assume connectivity in the population. Ideally, the target population is given by the union of the contact circles of all the persons in the initial sample. Otherwise, the initial sample suffers under-coverage, to which situation we will return later. The parameter λ can be envisaged as a product $M\psi$, where M is the size of the contact circle, and ψ is the probability of nomination. In this way, the Poisson distribution can arise as the limiting distribution of the Binomial distribution, denoted by $Bin(M, \psi)$, as M tends to infinity and ψ to zero and $M\psi$ to λ. Under the homogenous model, we assume that M and ψ are constants across the population. But heterogeneity can be introduced by allowing individual parameters $\lambda_i = M_i\psi_i$.

The conditional Binomial distribution (iii) of r_i given m_i corresponds to an assumption of *indiscrimination* when making the nominations. Since a person in the initial sample is not supposed to nominate him-/herself, the probability of the conditional Binomial distribution is $(n-1)/(N-1)$. Assumptions (ii) and (iii) are jointly equivalent to independent Poisson distributions of $r_i \sim Pois(\lambda_1)$ and $m_i - r_i \sim Pois(\lambda_2)$, where $\lambda_1/\lambda_2 = (n-1)/(N-n)$. This can be the case because, provided Bernoulli sampling (i) of the initial sample, the contact circle can fall either inside or outside of the initial sample, with expected $M_1 = M(n-1)/(N-1)$ and

$M_2 = M(N-n)/(N-1)$ persons, respectively. Indiscrimination of nomination, i.e. constant ψ in- or outside of s, yields then $\lambda_1/\lambda_2 = (M_1\psi)/(M_2\psi) = (n-1)/(N-n)$.

Under the homogenous sampling model, the likelihood of (N, p, λ) is given by

$$L(N,p,\lambda) \propto P(n,m,r;p,\lambda|N) = P(n;p|N)P(m;\lambda|n)P(r|m,n,N)$$

which admits a factorization between the likelihood of (N,p) and that of λ. It follows that λ is a nuisance parameter that can be ignored when the inference is focused on the size of the target population. The log-likelihood is given by

$$l(N,p) = \sum_{k=1}^{n} \log(N-n+k) + n\log p + (N-n)\log(1-p) - r\log(N-1) + (m-r)\log(1-(n-1)/(N-1))$$

For maximum likelihood estimation (MLE) we have $\hat{p} = n/N$, based on the equation $\partial l/\partial p = 0$. To find the MLE of N, we need to solve, on substitution of $\hat{p} = n/N$, the following equation

$$\sum_{k=1}^{n}(N-n+k)^{-1} + \log(1-p) - \frac{r}{N-1} + \frac{(m-r)(n-1)}{(N-1)(N-n)} = 0$$

This can be found by numerical iterations, denoted by $\hat{N}$.

For the numerical example cited above, we obtain $\hat{N} = 684$, practically equal to the conditional moment estimator under the graph model, i.e. $\tilde{N} = 685$. The homogenous sampling model seems to have captured the essence of the graph model of Frank and Snijders (1994).

C.3 Under-coverage of initial sampling frame

In the above we have assumed that every member of the target population has a non-zero probability of being included in the initial sample. If not, we say the initial sampling frame has under-coverage. Meanwhile, we say the single-stage LTS has under-coverage if the target population is larger than the union of the contact circles of the initial sample. Thus, the single-stage LTS may have full coverage, even when the initial sampling frame has under-coverage. Suppose, in such a situation, the initial sampling frame is a subset of the target population, denoted by U_1, of the size N_1 that is unknown. Assume that it is possible to distinguish the nominees who are (a) within the initial sample, whose count is given by r_{0i}, b) outside the sample but within U_1, whose count is given by r_{1i}, and (c) outside of U_1. Consider the following model.

(i) Assume Bernoulli sampling of the initial sample, denoted by $n \sim Bin(N_1, p)$.
(ii) Assume independent Poisson distribution of nominations m_i, denoted by $m_i \sim Pois(\lambda)$.
(iii) Conditional on m_i, assume the Multinomial distribution of $(r_{0i}, r_{1i}, m_i - r_{0i} - r_{1i})$ with respective probabilities $(\theta_0, \theta_1, \theta_2) = ((n-1)/(N-1), (N_1-n)/(N-1), 1-(N_1-1)/(N-1))$.

Again, assumption (ii) and (iii) are jointly equivalent to assuming independent Poisson distributions for $(r_{0i}, r_{1i}, m_i - r_{0i} - r_{1i})$ with parameters $(\lambda_0, \lambda_1, \lambda_2)$, where $\lambda_0 : \lambda_1 : \lambda_2 = (n-1) :$

$(N_1 - n) : (N - N_1)$. Let $r_0 = \sum_{i \in s} r_{0i}$ and $r_1 = \sum_{i \in s} r_{1i}$. The likelihood of (N_1, N, p) is given by

$$L(N_1, N, p) \propto P(n; p|N_1) P(r_0, r_1, m - r_0 - r_1 | m, n, N_1, N)$$

where the likelihood of λ factorizes away as above. The log-likelihood is given by

$$\begin{aligned} l(N_1, N, p) = & \sum_{k=1}^{n} \log(N_1 - n + k) + n \log p + (N_1 - p) \log(1 - p) \\ & - r_0 \log(N - 1) + r_1 \log(\frac{N_1 - n}{N - 1}) + (m - r_0 - r_1) \log(1 - \frac{N_1 - 1}{N - 1}) \end{aligned}$$

C.4 Stratified population

Sometimes stratification can help to avoid under-coverage of the initial sampling frame. For instance, one may divide the country into large police districts as strata. The data necessary for estimation can be created as follows. Suppose the target population is divided into a number of strata, denoted by N_h, for $h = 1, ..., H$, each with unknown size N_h. Suppose that it is possible to distinguish whether or not a person in the initial sample and a nominee by this person belong to the same *population* stratum. Denote by r_{hi} the number of nominees belong to U_h and denote by $m_{hi} - r_{hi}$ the number of nominees outside of U_h. Consider the following model.

(i) Assume stratified Bernoulli sampling of s, denoted by $n_h \sim Bin(N_h, p)$, for $h = 1, ..., H$.
(ii) Assume independent Poisson distribution of nominations m_{hi}, denoted by $m_{hi} \sim Pois(\lambda_h)$.
(iii) Conditional on m_{hi}, assume Binomial distribution of $r_{hi} \sim Bin(m_{hi}, (N_h - 1)/(N - 1))$.

Notice that the nominees are not required to be identified in the initial sample, but only within the population strata. Again, the likelihood of $\mathbf{N} = (N_1, ..., N_H)$, p and $\lambda = (\lambda_1, ..., \lambda_H)$ factorizes into that of $(\mathbf{N}, p)$ and that of λ. Let $r_h = \sum_{i=1}^{n_h} r_{hi}$. The log-likelihood is given by

$$\begin{aligned} l(\mathbf{N}, p) = & \sum_{h} \sum_{k=1}^{n_h} \log(N_h - n_h + k) + n \log p + (N - n) \log(1 - p) \\ & + \sum_{h} r_h \log(N_h / N) + \sum_{h} (m_h - r_h) \log(1 - N_h / N) \end{aligned}$$

References

Beverton, R.J.H. and Holt, S.J. (1957). On the dynamics of exploited fish populations. *Fish. Invest. Minist. Agric. Fish. Food G.B.*, Ser. II 19, 533p.

Félix-Medina, M.H. and Thompson, S.K. (2004). Combining link-tracing sampling and cluster sampling to estimate the size of hidden populations. *Journal of Official Statistics*, **20**, 19–38.

Frank, O. and Snijders, T. (1994). Estimating the size of hidden populations using snowball sampling. *Journal of Official Statistics*, **10**, 53–67.

Jandl, M. (2004). The estimation of illegal migration in Europe. *Studi Emigrazione/Migration Studies*, **XLI**(153), 141–155.

McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models*. London: Chapman and Hall.

Passel, J.S. (2007). *Unauthorized Migrants in the United States: Estimates, Methods and Characteristics*. Organisation for Economic Co-operation and Development, DELSA/ELSA/WP2(2007)2.

Pinkerton, C., McLaughlan, G., and Salt, J. (2004). Sizing the illegally resident population in the UK. Technical report, Migration Research Unit, University College London. Home Office Online Report 58/04, UK.

Thompson, S.K. and Frank, O. (2000). Model-based estimation with link-tracing sampling designs. *Survey Methodology*, **26**, 87–98.

Van der Heijden, P.G.M., Bustami, R., Cruyff, M., Engbersen, G., and Van Houwellingen, H.C. (2003). Point and interval estimation of the population size using the truncated Poisson regression model. *Statistical Modelling*, **3**, 305–322.

Van der Heijden, P.G.M., Cruyff, M., and Van Houwellingen, H.C. (2003). Estimating the size of a criminal population from police records using the truncated Poisson regression model. *Statistica Neerlandica*, **57**, 289–304.