Stockholm Research Reports in Demography | no 2018:24

Over-coverage in Swedish Population Registers Leads to Bias in Demographic Estimates

Andrea Monti, Sven Drefahl, Eleonora Mussino, Juho Härkönen

ISSN 2002-617X | Department of Sociology

Stockholm University
Demography Unit

Over-coverage in Swedish Population Registers Leads to Bias in Demographic Estimates

Andrea Monti[1 3], Sven Drefahl[1], Eleonora Mussino[1], Juho Härkönen[1 2]

[1] Demography Unit (SUDA), Department of Sociology, Stockholm University

[2] European University Institute (EUI), Department of Political and Social Sciences, Fiesole [3] Corresponding author: andrea.monti@sociology.su.se

Abstract: Estimating the number of individuals that live in a country has always been among the essential tasks for demographers. In this study we assess the potential bias in estimating the size of different migrant populations due to over-coverage in a country's population-register system. Over-coverage, i.e., from individuals registered but not living in a country, constitutes an increasingly pressing phenomenon and is tightly linked to differential patterns of registered emigration. However, there is no common understanding on how to deal with over-coverage in demographic estimates and research. This study examines different approaches to over-coverage estimation and discusses ways of improving current estimation methods using Swedish total population register data for the years 1990-2012. We assess over-coverage levels across migrant groups, test how estimates of age-specific fertility and mortality are affected when adjusting for over-coverage, and examine whether over-coverage can explain parts of the healthy migrant paradox. Our results confirm the existence of over-coverage. When adjusting for over-coverage, we find substantial changes in mortality and fertility rates for people in migrating ages. Our results suggest that accounting for over-coverage is particularly essential for correctly estimating fertility in migrant populations.

Keywords: Over-coverage, Fertility, Mortality, foreign-born, Sweden, register-based

Stockholm Research Reports in Demography 2018:24

ISSN 2002-617X

Introduction

Estimating population size is a fundamental task of demographers. Accurate estimation of population size is particularly challenging in the case of migrant populations and may lead both to underestimating and overestimating their size. Underestimations (under-coverage) of migrant populations are a problem in the case of unregistered immigrants (e.g. Woodrow and Passel 1990, Strozza 2004, van der Heijden et al. 2006). At the same time, many population registration systems lack accurate documentation of emigrations, due to lack of knowledge of the need to register an emigration or low incentives and compliance to do so, leading to over-coverage in population registers.

In addition to leading to inaccurate representations of the stocks of migrant populations and of their characteristics, the problems with estimating the size of these populations can contribute to biased estimates of core demographic phenomena and demographic rates. Even when (immigrant) populations are not accurately documented, their vital events (such as births and deaths) often are, leading to an inflation of the respective demographic rates in the case of under-coverage. Over-coverage leads to the opposite problem, in which vital events remain undocumented even though emigrated individuals continue to be erroneously regarded as being at risk of the event (e.g., Weitoft et al. 1999; Qvist 1999, Loeb et al. 2013). These problems can contribute to apparent demographic paradoxes, such as the unexpectedly low mortality rates and low fertility rates in some migrant populations (e.g. Qvist 1999; Palloni and Arias 2004).

In this study, we compare procedures to identify over-coverage in population registers with a focus on Sweden. Previous research has focused more on issues of under-coverage and the problem of estimating the size and characteristics of undocumented populations than on over-coverage. However, the problems associated with over-coverage can become more pertinent due to ongoing changes in demographic data collection, in which an increasing number of countries have moved to register-based systems and register-based censuses, as well as due to increases in migrant re-emigration (Castles et al. 2009; Jeffery & Murison 2011) and circular migration (Aradhya et al. 2017). Consequently, over-coverage has been identified as a potential source of bias for register-based censuses, official statistics, population forecasts, and academic research (e.g. Cortese and Greco 1993, Crescenzi et al. 2008, Crescenzi et al. 2009, Fortini et al. 2007, Statistics Sweden 2015a), but also for survey sampling (Commissione per la Garanzia dell'Informazione Statistica 2002; Martin et al. 2015; Salentin 2014) among migrants in particular (Maehler 2017).

Despite general acknowledgment of the problems of over-coverage, there is to date no praxis for

identifying the prevalence of over-coverage or for assessing its consequences for demographic research. We compare estimates of over-coverage based on different approaches suggested by researchers (Aradhya et al. 2017) and Statistics Sweden (SCB 2015a; Qvist 1999), both of which rely on traces of activity and the lack of such traces as reported in the register. More specifically, we estimate the prevalence of over-coverage and its trend and compare prevalence of over-coverage in different immigrant groups. Finally, we assess the potential bias from over-coverage on age-specific fertility and mortality rates. Sweden provides a good setting for this research given its comprehensive and widely-used population registration system, as well as its large and heterogeneous migrant population, which in relative terms is bigger than in the US[1], and its relatively high rates of re-emigration[2]. Although we focus on Sweden, the results carry lessons for estimating over-coverage and its consequences elsewhere as well.

Previous approaches

Over-coverage is a persisting issue for all countries with a sizable immigrant resident population, however, the extent of the problem and its variation across countries is largely unknown and directly related to different definitions of international long-term migrants in different countries. For the Swedish case, some earlier studies have attempted to address this issue for different migrant populations. For Finnish migrants in Sweden, Kirwan and Harrigan (1986) found an over-coverage rate of about 2.5 percent and concluded that an error of that magnitude is unlikely to bias conclusions for their studied outcomes. They, however, had the advantage to access both Swedish and Finnish data and they caution that departures to non- Nordic destinations may be more problematic to address (Kirwan and Harrigan 1986). Statistics Sweden also addressed the same issue in a report in the late 1990s, concluding that over-coverage of immigrants recorded as residing in Sweden is about 1 percent for migrants from Nordic countries and 2.8 percent for migrants from other countries (Qvist 1999). In as recent paper, Ludvigsson et al. (2016) conclude from personal communications with Statistics Sweden that their estimate for over-coverage is equal to 0.25–0.5% of the entire Swedish population. The over-coverage of Nordic immigrants may be about 0.1%, but substantially higher for individuals born outside the Nordic countries (potentially 4–8 percent). Statistics Sweden also argues that the often low mortality in foreign-

[1] *In the end of 2016, Sweden had 17.85% foreign born residents registered in the total population (Statistics Sweden, 2018). According to the United States Census Bureau (2018), the equivalent share for the US the same year was 13.25%*

[2] *Among migrants migrating to Sweden between 1990 and 1995 almost 27% had emigrated from Sweden within 10 to 15 years (Monti 2018).*

born individuals suggest that a significant proportion of them no longer reside in Sweden, with substantial variation for country of origin and age (Statistics Sweden 2015a). The impact of over-coverage for the estimation of demographic rates is thought to be largest at the very highest ages, where the number of live individuals becomes smaller and registration errors tend to accumulate (Statistics Sweden 2015a). Consequently, the Swedish Tax Agency performs routine checks on individuals aged 100 and above.

Previous studies (e.g., Syse et al. 2016 for Norway; Wallace and Kulu 2014 for England and Wales; and Turra and Elo 2008 for US) applied different "correction methods" to explain the lower mortality among migrants vs their host populations (the healthy migrant paradox). However, nobody found a reliable and repeatable measure. Recently, Aradhya et al. (2017) suggested to deal with over-coverage in register-based research using an income-based exclusion method. The suggestion is based on the idea that all individuals without any economic activity in a welfare state like Sweden in a given year can be assumed to not live in the country and thus should be excluded from the study population. Using this criterion, which we call the *zero personal income* approach in the remainder of this study, is a relatively straightforward way to exclude individuals who are thought to not belong to the population counts (any longer). While this solution is appealing because zero-income individuals can easily be identified in most register-based research, there is very little known about its appropriateness. A second approach, which we will call the *register-trace* approach in the remainder of this paper, has been proposed by Statistics Sweden in their efforts to evaluate the quality of the population registers. This approach tracks a larger number of activities in different linked Swedish registers (Statistics Sweden 2015a). In 2015, Statistics Sweden further developed the *register-trace* approach, considering not only cross-sectional but also longitudinal information. One of the novelties of this study is to compare and examine the different ways of over-coverage estimation discussed above and to show the impact of these over-coverage measures on demographic estimates of fertility and mortality.

Data and Method

The data used for this study are Swedish administrative register data on foreign-born residents in Sweden aged 18-75 years old during 1990 to 2012, who have been registered in the official national population register of the total population. Detailed annual data are derived from several administrative registers[3], and enable us to create different measures of over-coverage.

[3] *Included are registers on the Total Population, Social Insurance, Emigration and Immigration, Domestic migration, Cause of Death, Civil Status Changes, and Education.*

In Sweden, individuals whose main place of actual or planned residence[4] is within the country for at least one year are registered in the official national population registers. The incentives to become registered are high since basically all formal contact with authorities and other institutions require individuals to be so. If leaving the country for at least one year, individuals are obliged to report their emigration and thus be de-registered. However, the incentives to do so are low and knowledge about this obligation is limited.

To empirically detect over-coverage, different attempts have focused on confirming individual presence by looking at officially recognized activities. It is reasoned that if a person indeed resides in Sweden, this should be visible in national registers somehow. Following previous studies we replicate[5] and compare three different ways of validating presence in Sweden by searching for activity in the registers. The individuals not found active by any of these approaches will be defined as contributing to the over-coverage. Each approach is described below.

- The *zero personal income* approach

  One way to ensure correct coverage of study populations in empirical studies has been to exclude people with no personal income (i.e., Aradhya et al. 2017; Weitoft et al. 1999). The argument is that with no economic means to secure one's livelihood, it is unlikely that a person is regularly active in the country. In this approach, individuals are classified as not residing in Sweden (over-covered) in a given year if he or she has no reported personal income from earnings, social allowances, parental leave, sick leave, student financing, unemployment benefits, labor market programs, elderly pensions, home care allowances and other pensions and social benefits.

  This intuitively appealing approach requires access to a dozen or so variables that are routinely available in register-based data. Although it is likely to correctly classify over-coverage in a large share of cases, a limitation of this approach is that it does not apply to children and youth. Additionally, there is a risk of excluding residing people who for some reason do not have any registered income, for example through uncertain employments on the black market or through family support. On the other side, it is still possible to have a positive income while living

[4] *"Residing" in Sweden requires spending your daily rest in the country on a regular basis, corresponding to at least 52 days a year (SFS 1991:481).*

[5] *Our register-trace variables are similar but not identical to the ones Statistics Sweden uses (2015a). Differences are due to not accessing exactly the same register variables and also the fact that we are looking at a specific age span of 18-75 years. For example "being born" is therefore not part of our measure.*

outside the country, for example through pensions or Swedish employment located abroad.

- The *register-trace* approaches

A broader approach is used by Statistics Sweden in order to assess national register quality. It is based on the same logic, which is that regularly residing individuals should show some type of activity in the national registers. In this "*register-trace*" approach, personal income is only one of several activities that would vouch for individual presence. Included are also vital events, household income and educational changes. We employ two versions of the *register-trace* approach, a *cross-sectional* version that is only based on information during a single calendar year, and a longitudinal version based on information of three subsequent years. In the cross-sectional *register-trace* approach, over-coverage is assumed when a person is not found active in any of the following domains in a given year[6]:

- Immigration
- Emigration
- Change of civil status (though not due to the death of spouse)
- Change of citizenship
- Domestic move within the country
- Graduation from the gymnasium
- Enrollment in any higher education (from gymnasium level and above), measured both from information on student allowance and latest year of obtaining course credits
- Employment (including self-employment if reaching a certain level of income)
- Unemployment or in unemployment program, as registered by the Public Employment Service
- Being linked to any household income, measured as the sum of the personal incomes of all members in a household
- Death

The longitudinal *register-trace* approach extends on this idea in that non-activity (in the

[6] *Over-coverage is by all three approaches defined when a person is registered in the national population register but not active once during the same year. This means that a person living in Sweden in January, leaving the country in February and registered as part of the Swedish population in November, will still not be considered as contributing to the over-coverage. Our measures of over-coverage are thereby probably slightly underestimated compared to the same approach but with all calculations made on monthly data.*

specific year for which over-coverage is estimated) might be more or less plausible given any past and future circumstances. In 2015, Statistics Sweden developed its longitudinal *register-trace* approach allowing also for past and subsequent register activity. This longitudinal approach applies to individuals not found active according to the cross-sectional *register-trace*. It is based on a sum of weighted indicators considering activity one year prior and one year past the given year in relation to individual characteristics. The different register-based indicators are summarized within two groups, one indicating correctly registered residence and another indicating over-coverage. The specific indicators and their weights used in this study all originate from the method proposed by Statistics Sweden (2015a), and are summarized in Table 1 in the Appendix. If indicators signifying correct registration (indicators 1-6 in Table 1 of the Appendix) exceed the indicators signifying over-coverage (indicators 7-18), the non-active individuals from the cross-sectional approach are no longer considered part of the over-coverage. This means that less people are marked as contributing to the over-coverage than according to the cross-sectional *register-trace*.

The advantage of the *register-trace* approaches is that they cover a larger number of life domains and thus should be able to identify activity also for those individuals without any own economic activity. That said, the approach clearly has much higher data requirements, particularly if longitudinal information is included.

We estimate over-coverage according to all three approaches for each calendar year during 1990-2012. Using the reference year of 2010[7] we compare observed and adjusted age specific death and fertility rates (ASFR and ASDR) in the studied populations. Following our results, we analyze the consequences of overestimating the foreign born population in demographic research. Based on the comparison of different measures we propose which of the indicators used in the *register-trace* approach that best complement the *zero personal income* approach as an adjustment of over-coverage. The aim is to derive an estimation that is similar to existing *register-trace* approaches with a more parsimonious combination of indicators.

The adjustment for over-coverage in a particular year can be made by excluding all individuals who according to the different approaches are characterized as being part of the over-coverage from the studied population. Evidently, most adjustments relate to denominator data. Thus, this can often be done by solely excluding the over-covered individuals from contributing to risk time,

[7] *2010 is the last year allowing us to properly define also the longitudinal register approach.*

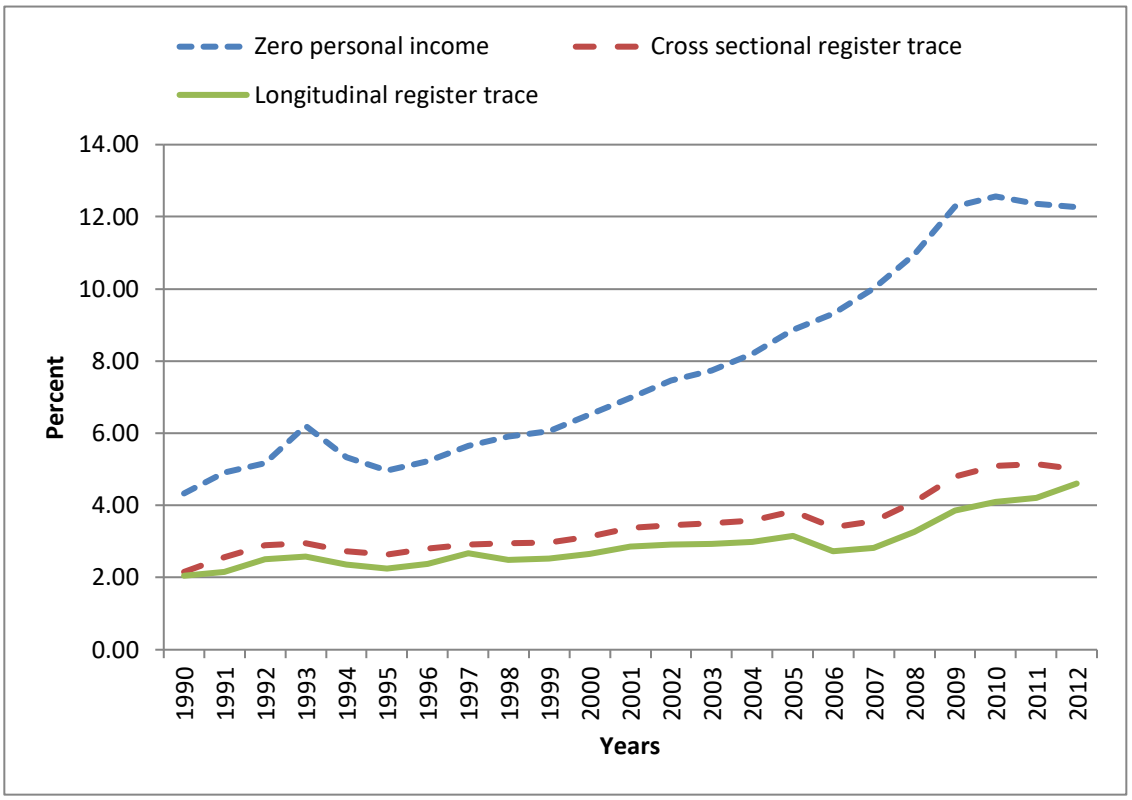
i.e., from the denominator. Similar to Aradhya et al. (2017), for the personal income approach, we choose to exclude over-covered individuals both in any nominator and denominator data when calculating corresponding demographic rates. The difference is minor since one of our observed events (death) is already part of the two *register-trace* approaches.

Because over-coverage in Sweden is a phenomenon mostly linked to foreign-born migrants (Statistics Sweden 2015a) in this study we focus on foreign-born population.

Results

Over-coverage in Sweden increased according to all three indicators during 1990–2012. **Figure 1** shows the proportion of individuals that are assumed to contribute to the over-coverage based on the three different estimation approaches. Comparing the different approaches of over-coverage measurement shows how the two *register-trace* approaches give lower and more similar estimates and are more stable over time, whilst the zero-income approach produces higher shares of over-coverage over the whole period and is also more volatile. Estimates of the prevalence of over-coverage range from around 4% of the foreign-born population according to the *register-trace* measures to up to 12% when using the zero-income approach. For example, estimates of over-coverage increased in the beginning of the 1990's, when Sweden received large numbers of Yugoslavian migrants, who during the first year(s) didn't receive any personal income of the kind captured by our income variable. The overall increase over time in over-coverage estimation can also be related to the overall rise in registered (and non-registered) emigration during the same period, a trend that has been especially noted among the foreign-born but is also prevalent for the Swedish-born population (Statistics Sweden 2015b).

Figure 1: Over-coverage among foreign-born residents according to different approaches, 1990–2012.



Source: Swedish register data. Calculations made by the authors.

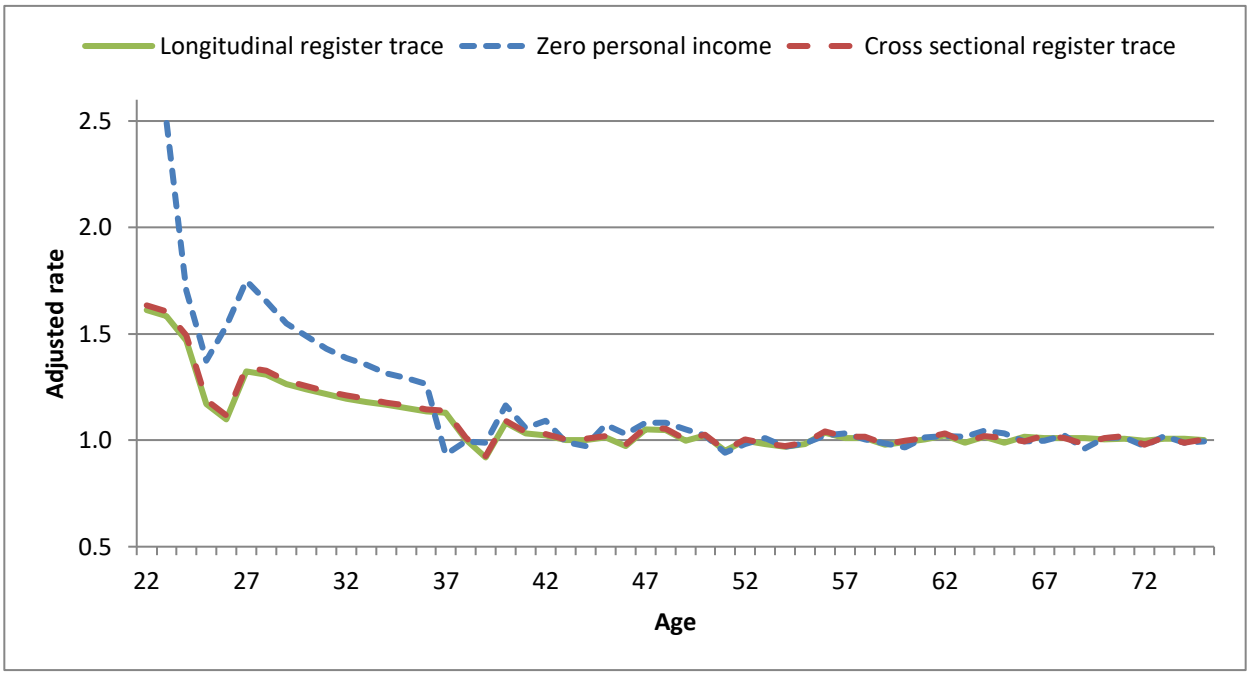
In the next step we examine consequences of over-coverage for the estimation of demographic rates. To show the possible impact of over-coverage we estimated fertility and mortality rates for 2010, before and after adjusting for over-coverage. We hypothesize that the bias varies by age and that it is sensitive to the type of process we study, the age-specific intensity of that process, and the age-specific intensity of migration.

Figure 2 shows the relative differences in mortality rates between mortality rates with and without adjustment for over-coverage for ages 18-75. Three different measures of over-coverage are presented. A value of 1 would indicate that there is no difference in the mortality rate before and after adjustment for over-coverage, suggesting that the impact of over-coverage on mortality rates is negligible. However, if over-coverage is present, adjusted rates should be higher, as individuals are removed from the denominators of the rate calculations. In terms of mortality we find a large impact of over-coverage adjustment for ages with high migration intensity, up to around age 40, and low to very low impact for ages 40 and above. When using the zero-income approach mortality rates are up to 2.5 times higher after adjustment at ages 20-30. When using the longitudinal and cross-sectional *register-trace* approach we find mortality rate differences of about 25-50% at those ages. With our data we are not able to address the impact of over-coverage

for ages of high mortality intensity, that is after age 75. While the observed patterns do not suggest that the impact of over-coverage increases again at higher ages.

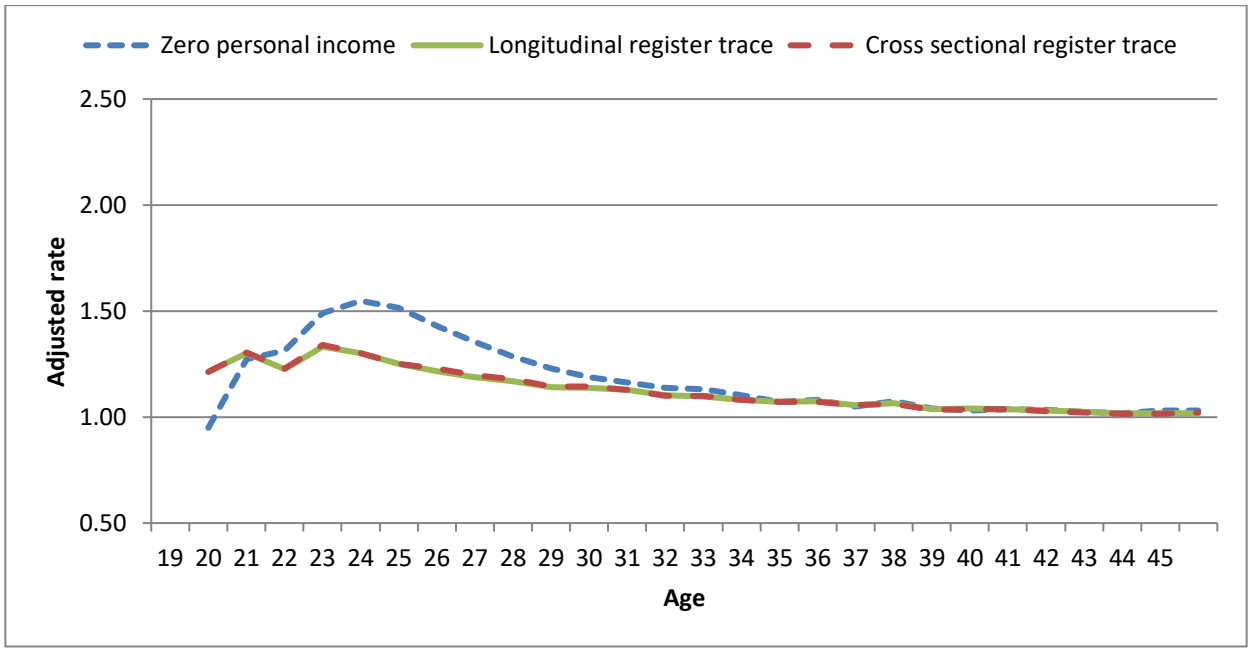
Figure 3 gives the corresponding results for estimates of fertility rates. Similar as for mortality, adjusting for over-coverage increases rates, particularly if we use the *zero personal income* approach. Using this indicator, we find a peak in the rate difference, meaning that fertility rates are about 50% higher after adjustment, at around age 24. Using the longitudinal and cross-sectional *register-trace* we observe less dramatic but still very substantive estimate differences for many ages with high fertility intensity, with an emphasis towards the younger part of the fertility schedule. The differences between the indicators are highest for ages where women are more likely to be out of the labor market, e.g., when they are studying, and thus have a higher risk of having zero observed personal income.

Figure 2: Age Specific Death Rates among the foreign-born population in Sweden, adjusted for over-coverage as a ratio of observed values, year 2010.



The table shows the ratio of adjusted ASDR divided by the observed values (value 1). Source: Swedish register data. Calculations made by the authors.

Figure 3: Age-specific fertility rates among the foreign-born population in Sweden, adjusted for over-coverage as a ratio of observed values, year 2010.

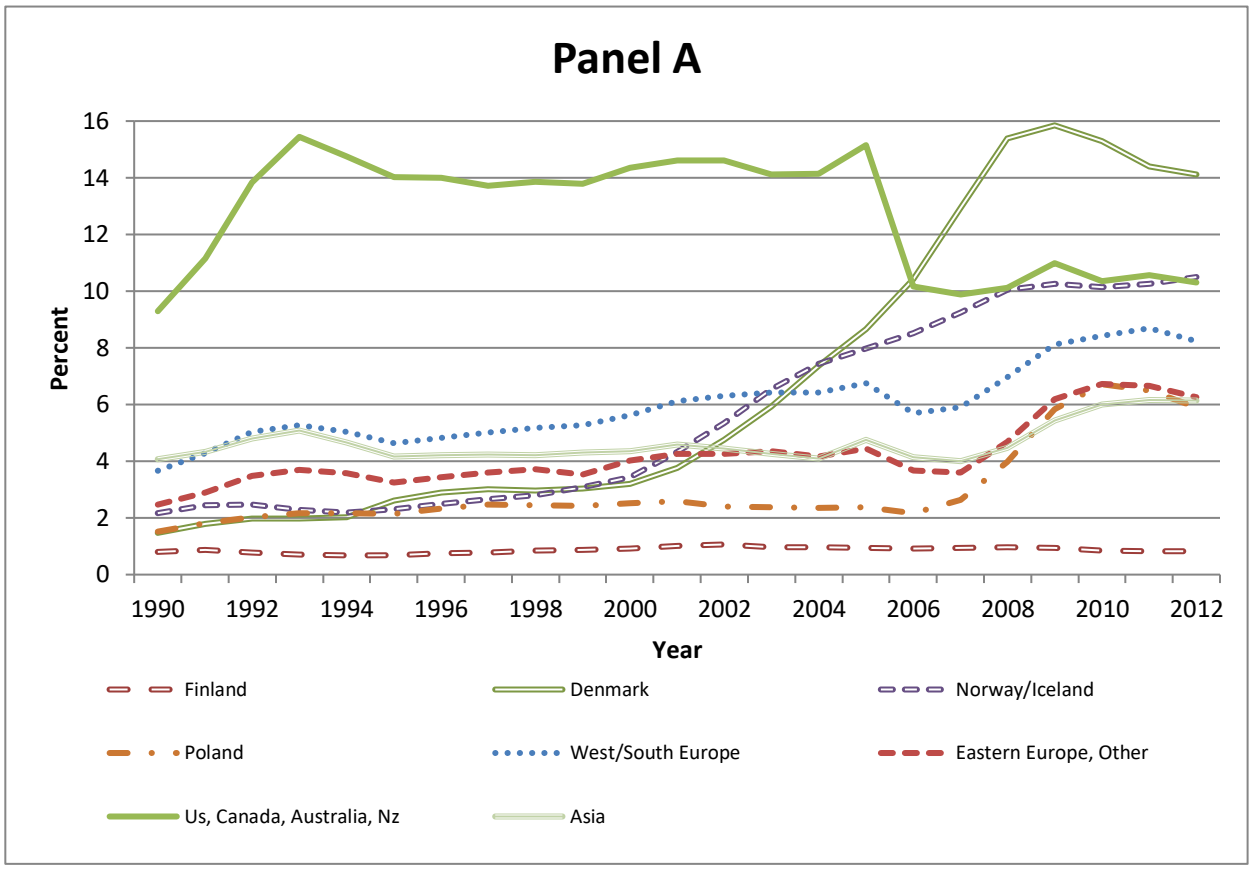


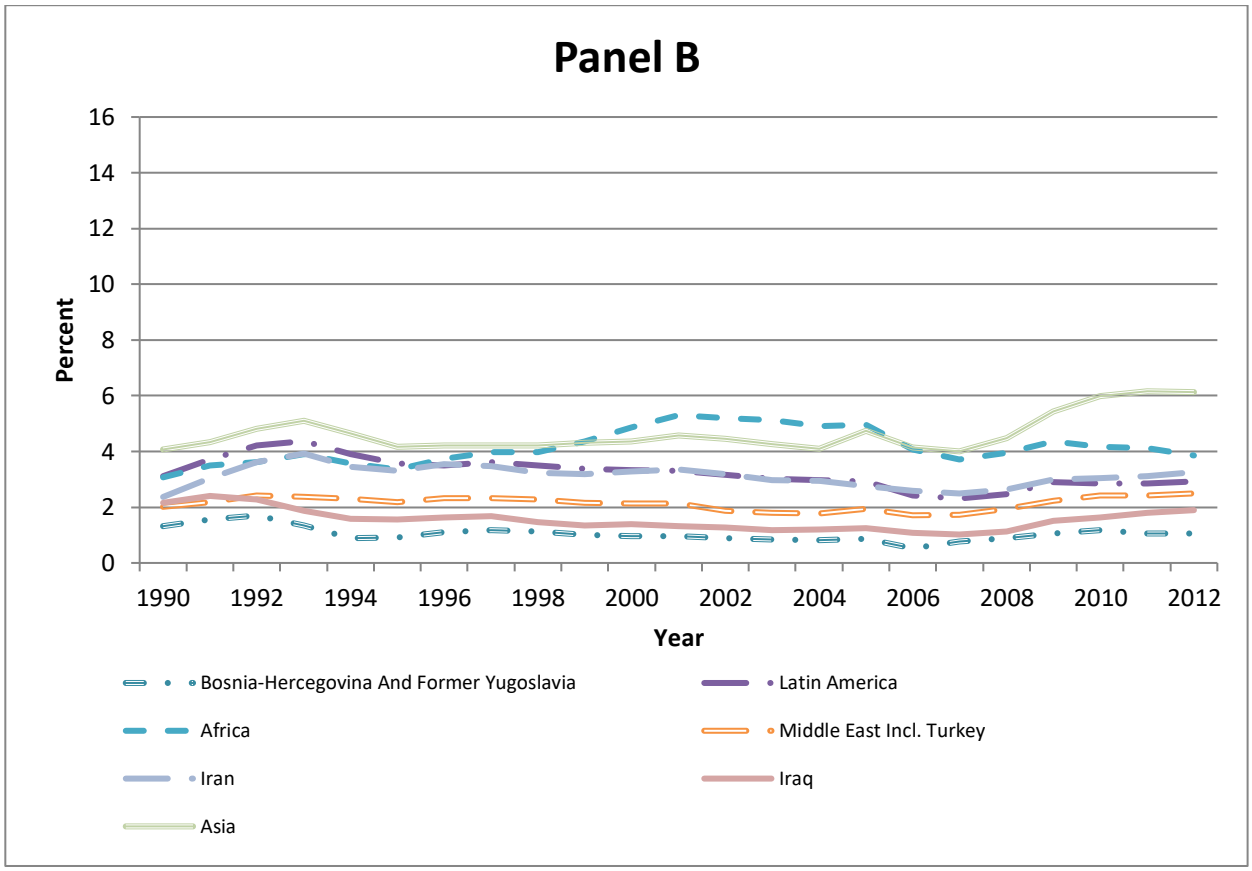
The table shows the ratio of adjusted ASFR divided by the observed values (value 1). Source: Swedish register data. Calculations made by the authors.

For both, mortality and fertility, over-coverage seems not to lead to a substantive bias for ages with low migration intensity, i.e., at ages after the mid-30s. For younger ages, the bias is important for both processes. However, its real-world impact is likely more substantive for fertility estimates as the bias is concentrated in ages of high fertility but low mortality. If the bias is high at ages of high process intensity, both relative and absolute differences become high, meaning that projection estimates of the numbers of births would be more biased than that of the number of deaths.

So far, the results have shown differences between the *zero personal income* and the *register-trace* approaches that tend to increase over time both in relative and absolute terms. As mentioned, *zero personal income* is sensitive to societal changes and only using personal income as exclusion criteria runs a much higher risk of excluding individuals that actually belong to the population. However, few differences are observed between the cross-sectional and longitudinal *register-trace* approaches. For that reason, when looking at the difference of over-coverage by country of origin we decided to show the results based on the cross-sectional *register-trace* indicator, as the best-quality and most-easy to use indicator at hand (**Figure 4**, Panel A and B).

Figure 4: Over-coverage in percent (measured through cross-sectional *register-traces*), by country of birth, years 1990–2012.





Source: Swedish register data. Calculations made by the authors.

Between 1990–2012 the overall differences in over-coverage levels by country of birth have increased. To large extent these differences can be explained by variations in registered emigration across the groups (Appendix **Figure 6**, Panels A and B). For example, migrants born in the US, Canada, Australia or New Zealand exhibit the highest proportion of over-coverage during the overall period, which correspond to their high emigration rates. In 2005, the Swedish Tax Authority conducted a larger control of possible over-coverage and corrected the number of people assumed to be living or not in the country (Swedish Tax Authorities 2006), which could explain the sharp decline in over-coverage for migrants from some countries in 2006 (**Figure 4**, Panel A). People not found active in the registers were registered as having emigrated in 2005, which led to a sharp increase in emigration numbers for these country groups (Appendix **Figure 6Figure 6**, Panel A).

The expansion of EU member countries and more countries joining the Schengen-agreement in 2007 made it easier to move within EU-member states. As a consequence, we observe a notable increase in over-coverage rates during this time, especially for Western, South and Eastern European countries, among them Poland, but also from Asia, Latin America and countries like Iran, Iraq and other Middle Eastern countries. A similar but smaller increase is noted for the previous EU expansion in 2004.

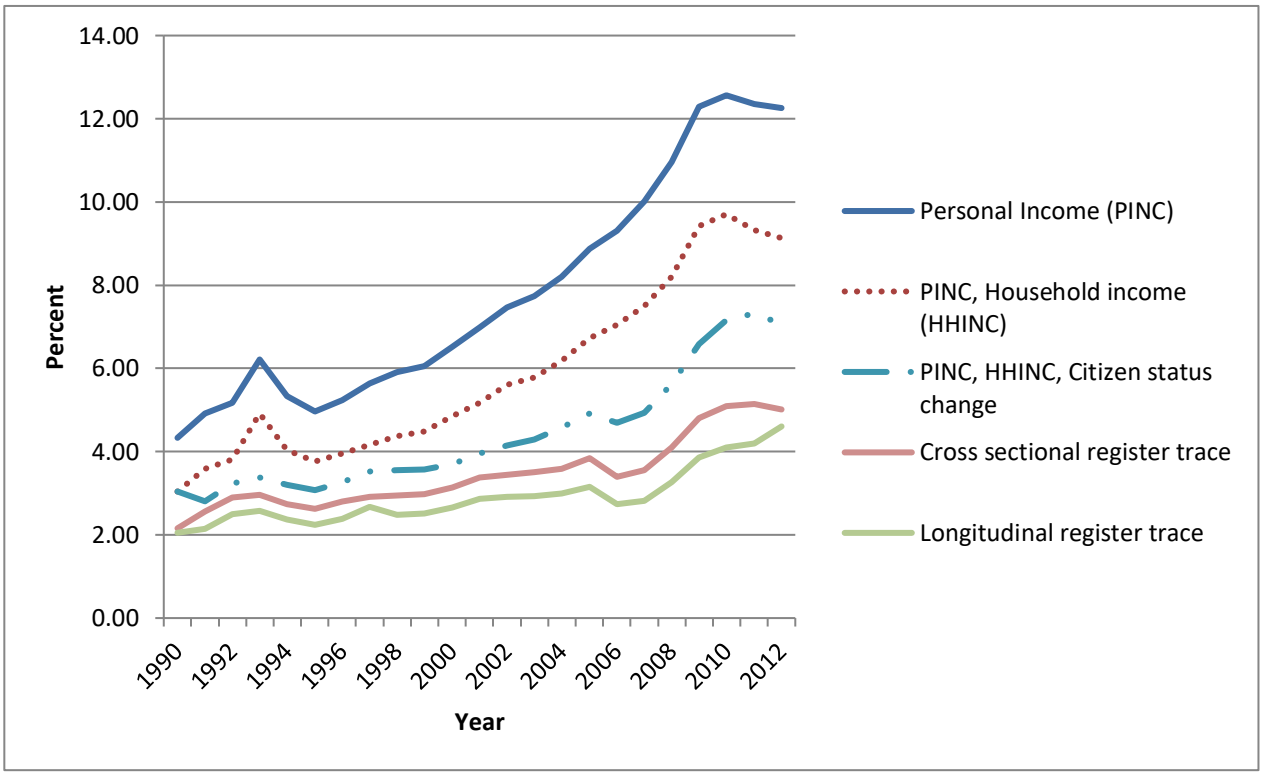
Due to a Nordic agreement all intra-Nordic immigration is automatically reported to the sending country. This agreement can partly explain both the low over coverage rates for Finnish born migrants and the high rates for Danish and Norwegian migrants. As soon as an individual registers its presence in a Nordic country, for example a Finnish-born migrant in Sweden who is now returning to Finland, a message is sent to, in this case Sweden, for de-registration. However, in regions with close geographic proximity, this system still leads to a larger number of false registrations. This is because individuals can work and live simultaneously in two countries and often register their presence in order to maximize their own economic advantages. In 2006, the Swedish Tax Authorities (2006) voiced its concern over practices of "false immigrations", as it contributes to over-coverage. In the case of Danish migrants, the opening of the Öresund-bridge between Sweden and Denmark in 2000 closely connected the densely populated areas of Malmö and Copenhagen, which have led to a continued increase in over-coverage for Danish nationals in Sweden.

Our study results have revealed big differences in the estimation of over-coverage using the *zero personal income* approach and the *register-trace* approach, with *zero personal income* overestimating over-coverage to a very large extent. However, the fact remains that the full

register-tracing of individuals requires a lot of information that the analyst seldom has easily at hand, whilst personal income is a more easily accessible variable. In the next step we aim to find an accommodating way of improving current estimation methods by adding one variable at a time from the cross-sectional *register-trace* approach to the *zero personal income* approach. The purpose is to derive a skimmed-down version of the register-trace approach that is easier to apply than the full version. From this procedure, the variable that seems to discriminate most, together with the *personal income*, is *household income*. For graphical reason all the steps of this procedure are not reported but in **Figure 5** we show the most parsimonious combinations of added variables to the personal-income approach.

Adding a second variable to the *zero personal* plus household income approach reduces the gap to the full *register-trace* approach even more. *Zero personal income* should thereby not be rejected as a basis for over-coverage but should be complemented with additional information. Adding another variable or two means that much less people are counted as contributing to the over-coverage (**Figure 5**). Adding a variable on citizen status change decreases the difference between the two approaches even further. This combination should therefore be considered as the most preferable option.

Figure 5: Changes in over-coverage levels during 1990-2012 when adding single variables to the *zero personal income* approach



Source: Swedish register data. Calculations made by the authors.

Discussion

A number of apparent paradoxes in research on migrant populations stem from two sources of potential bias: 1) when migration movements are associated with the outcomes of interest, and 2) when migration movements are not recorded perfectly. In the latter case, paradoxes can occur due to the fact that some people are incorrectly classified as (not) being in the country. For example, administrative registers may not include all people that presently reside in a country, such as unregistered immigrants whose immigration events where not recorded. This may lead to an underestimation of the size of the foreign-born population, labelled as under-coverage. Much research has focused on issues related to under-coverage and the problem of estimating the size and characteristics of undocumented populations in a given host country. In this paper we address a closely associated bias, that of over-coverage, which may occur when emigration movements are not recorded correctly. Over-coverage has become more important over time because it is tightly linked to 1) migration processes, e.g., increases of re-emigration among migrants (Castles et al. 2009; Jeffery & Murison 2011) and 2) ongoing changes in demographic data collection, such as the number of countries that have moved to register-based data-collection systems and register-based censuses. In this study, we compared different indicators to assess over-coverage, the variation of over-coverage across migrant groups, and calculated the bias in demographic rates that is produced by over-coverage. Our focus has been on Sweden, a country with comprehensive population registration and a large and heterogeneous migrant population.

First, we evaluated a proposed *zero personal income* approach and addressed how to improve its accuracy. Then, we compared different approaches of over-coverage in terms of the estimated prevalence of over-coverage. By constructing two different versions of a *register-trace* approach we compared these with the *zero personal income* approach over time. We found that using solely *zero personal income* will likely overestimate over-coverage to a very large extent as compared to the *register-trace* approaches and that the differences between the approaches have increased over time.

Using different ways of estimating over-coverage we analyzed the extent to which over-coverage may bias estimates of mortality and fertility rates among immigrants in Sweden. Our results show that there is an upward adjustment of mortality rates among immigrant groups at ages of high migration when controlling for over-coverage, independently of what measure was used. This suggests that over-coverage could explain parts of the healthy migrant paradox, at least at the ages with high migration. The impact of over-coverage may be even more important for the correct estimation of fertility. According to our measures, any potential over-coverage bias is largest at

ages with relatively high fertility intensity, potentially underestimating age-specific fertility rates for women in their 20s by some 30-50%. This suggests that accounting for over-coverage is essential for correctly estimating fertility in migrant populations.

Previous reports (Statistics Sweden 2015a) have concluded that there are differences in over-coverage across different migration origins. We confirm large variations in over-coverage depending on the origin of migrant groups, as defined by their country of birth and also found that elevated over-coverage rates of specific migrant groups are associated with high emigration rates and the possibilities of free mobility. Therefore, it is particularly important to adjust for over-coverage in demographic studies of migrant populations with known high emigration intensities.

In order to improve on currently available estimation methods, we derived a parsimonious approach to see how adding selected variables to the *zero personal income* approach attenuated the prevalence estimation of over-coverage. Based on our analysis, we advise future users of the *zero personal income* approach to combine that variable with at least one additional measure of activity, preferably that of household income, in order to increase its accuracy. In order to get an even more stable measure to cover true people with no own income, an indicator of citizen status change could be added as a third variable.

Our results show that using both individual and household income increases the chances of a more accurate over-coverage estimation at the national level in Sweden. In specific study populations, for example with larger shares of single households, the improvements are likely less substantial and it might be advisable to combine the *zero-income* approach with additional measures. As an example we may consider regions close to the national border, sailors, long distance commuters or other transnationally mobile people, e.g., the individuals living in the Öresund region close to Denmark. Some of these people might be registered in Sweden and have their income based in Sweden, although they spend most of their day and night time on the other side of the bridge. How they are classified depends on what information happens to be entered into the registers of Sweden and Denmark and the choice of estimating method. Additionally, the correction using the zero personal income might introduce even more bias in other countries than the Nordics, where labor-force participation is much lower, and where social transfers are also less all-encompassing.

In sum, our results have shown that the impact of over-coverage can be substantial and that there may be biases in estimates of different measures on migrant populations also in countries with the highest quality of their registration systems. As such, our results carry lessons for estimating over-coverage and its consequences elsewhere as well. Research needs to acknowledge that any

demographic estimates based on migrant populations are likely to be biased at ages of high migration intensity and that currently available correction methods need to be improved further.

Our study was limited in that we could not access the impact of over-coverage above ages 75 because our data on many socioeconomic characteristics was limited to ages below 75. Future research needs to address this shortcoming because errors in the registration system are known to accumulate at older ages when population sizes tend to become so small that even a small number of errors in the vital registration system can lead to substantive bias of population level estimates.

Assessing error in register-based civil registration systems is becoming increasingly important now that numerous countries are moving from a traditional census to a register-based census. One of the primary goals of the traditional census was correcting population counts both in term of under- and over- coverage of the resident population. For Sweden, a country characterized by a long history of highest-quality registers, we show that the *register-trace* approach might help to reduce estimation errors at least in terms of over-coverage.

This study addressed bias in demographic estimates on migrant populations that stems from errors in data collection, in our case the correct recording of emigrations. However, there is at least one further and related source of bias that has been brought forward in migration research, which can occur when migration movements are associated with the outcomes that are studied. This may happen when demographic events occur either more or less frequently in periods that cannot be observed, e.g., before immigration or after emigration. Conclusions that are drawn from time periods with available data are likely systematically biased. The fertility literature points out one common example. Studies often find high fertility levels shortly after arrival in a destination country, which are produced when individuals postpone their fertility decision in anticipation of migration (e.g., Andersson 2004, Mussino and Strozza 2012, Milewski 2010). In turn, the mortality literature discusses the role of salmon effects in mortality, which are produced when individuals emigrate or return to their country of origin in case of poor health and anticipation of death. This would yield reduced mortality rates for immigrants in a given destination country (Abraido-Lanza et al. 1999, Andersson and Drefahl 2016). These examples are not based on actual problems of data coverage, they are rather manifestations of the endogeneity of migration and other demographic events, which can produce apparent paradoxes in the study of demographic processes. We argue that in order to improve our demographic and other estimates for migrant populations, researchers need to systematically distinguish and address all sources of potential bias.

Acknowledgements

We thank Gunnar Andersson for his valuable and constructive comments on the first draft of this manuscript. This research was supported by the Strategic Research Council of the Academy of Finland (TITA project, decision number: 293103); the Swedish Initiative for Research on Microdata in Social Science and Medical Sciences (SIMSAM): Stockholm University SIMSAM Node for Demographic Research, grant 340-2013-5164; and the Swedish Research Council for Health, Working life and Welfare (FORTE), grant 2016-07105.

References

Aradhya S., Scott K. & Smith C. (2017). Repeat immigration: A previously unobserved source of heterogeneity? Scandinavian Journal of Public Health, 45(Suppl 17), 25-29.

Abraído-Lanza A. F., Dohrenwend B. P., Ng-Mak D. S. & Turner J. B. (1999). The Latino Mortality Paradox: A Test of the "Salmon Bias" and Healthy Migrant Hypotheses. American Journal of Public Health, 89(10), 1543-1548.

Andersson G. (2004). Childbearing after migration: Fertility patterns of foreign-born women in Sweden. International Migration Review, 38(2), 747-774.

Andersson G. & Drefahl S. (2017). Long-Distance Migration and Mortality in Sweden: Testing the Salmon Bias and Healthy Migrant Hypotheses. Population Space and Place, 23(4). doi: 10.1002/psp.2032

Castles S., de Haas H. & Miller M. J. (2009). The Age of Migration: International Population Movements in the Modern World. Basingstoke: Palgrave Macmillan.

Commissione per la Garanzia dell'Informazione Statistica (2002). Il campionamento da liste anagrafiche: analisi degli effetti della qualità della base di campionamento sui risultati delle indagini, Rapporto di ricerca CGIS, n. 02.12, Roma Dicembre 2002

Cortese A. & Greco M. (1993). Il grado di copertura del censimento demografico 1991: considerazioni sulla base del confronto con le risultanze anagrafiche, Quaderni di Ricerca Istat, Roma 1993, Serie Interventi e Relazioni.

Crescenzi F., Fortini M., Gallo G. & Mancini A. (2008). Nota per il Presidente e il Consiglio dell'Istat, Linee generali di impostazione metodologica, tecnica e organizzativa del 15° Censimento generale della popolazione. Roma, settembre 2008.

Crescenzi F., Fortini M., Gallo G. & Mancini A. (2009). La progettazione dei censimenti generali 2010 – 2011 in Linee generali di impostazione metodologica, tecnica e organizzativa del 15° Censimento generale della popolazione. Documenti ISTAT, n. n. 6/2009

Maehler D. B., Martin S. & Rammstedt B. (2017). Coverage of the migrant population in large‐scale assessment surveys. Experiences from PIAAC in Germany. Large-scale Assessments in Education An IEA-ETS Research Institute Journal 20175:9 doi:10.1186/s40536-017-0044-8

Fortini M., Gallo G., Paluzzi E., Reale A. & Silvestrini A. (2007). Criticità di processo e di prodotto nel 14° Censimento generale della popolazione e delle abitazioni: aspetti rilevanti per la progettazione del 15° Censimento. In La progettazione dei censimenti generali 2010–2011, Documenti ISTAT, n. 10/2007

Frank R. & Heuveline P. (2005). A crossover in Mexican and Mexican-American fertility rates: evidence and explanations for an emerging paradox. Demographic Research, 12(4), 77-

Kirwan F. & Harrigan F. (1986). Swedish-Finnish Return Migration, Extent, Timing, And Information Flows, Demography, 23(3)

Jeffery L. & Murison J. (2011). The temporal, social, spatial, and legal dimensions of return and onward migration. Population, Space and Place, 17(2), 131–139.

Ludvigsson J. F., Almqvist C., Bonamy A. K., Ljung R., Michaelsson K., Neovius M., . . . & Ye W. (2016). Registers of the Swedish total population and their use in medical research. Eur J Epidemiol, 31(2), 125-136. doi:10.1007/s10654-016-0117-y

Milewski N. (2010) Fertility of immigrants: A two-generational approach in Germany. Heidelberg, Dordrecht, London, New York: Springer.

Monti A. (2018). Re-emigration of Foreign-Born Residents from Sweden: 1990-2015 Working Paper. Stockholm Research Reports in Demography, 2018:15, doi:10.17045/sthlmuni.6217247.v1

Mussino E. & Strozza S. (2012). The fertility of foreign immigrants after their arrival: The Italian case. Demographic Research, 26(4), 99-130.

Palloni A. & Arias E. (2004). Paradox lost: explaining the Hispanic adult mortality advantage. Demography, 41: 385-415.

Qvist J. (1999). Täckningsproblem i Registret over totalbefolkningenRTB---Skattning av övertäckning med en indirekt metod [Problems of coverage in the Register of Total Population (RTB)—estimation of overcoverage by an indirect method]. R & D Report. Stockholm, Sweden: Statistics Sweden; 1999

Salentin K. (2014). Sampling the ethnic minority population in Germany. The background to “migration background”. Methods, Data, Analyses, 8(1), 25–52, doi:10.12758/mda.2014.002

Strozza S. (2004). Estimates of the Illegal Foreigners in Italy: A Review of the Literature. International Migration Review, 38(1), 309–331.

Loeb S., Drevin L., Robinson D., Holmberg E., Carlsson S., Lambe M. & Stattin P. (2013). Risk of localized and advanced prostate cancer among immigrants versus native-born Swedish men: a nation-wide population-based study. Cancer Causes Control, 24(2), 383-390. doi:10.1007/s10552-012-0124-6

Statistics Sweden (2018). Swedish and foreign-born population by region, age and sex. Year 2000 - 2016 [Table.] Retrieved 2018-01-29

Statistics Sweden (2015a). Overcoverage in the Total Population Register—a register study. Background facts. Population and Welfare 2015;1. (Swedish: Statistiska Centralbyrån. Övertäckning i Registret över totalbefolkningen – en registerstudie. Bakgrundsfakta. Befolkning och Välfärd 2015:1.)

Statistics Sweden (2015b). Statistics on immigration and emigration. [www.scb.se]

Swedish Tax Authorities (2006). Mapping of the population registration error. Solna, Swedish Tax Authorities 2006:7. (Swedish: Skatteverket. Kartläggning av folkbokföringsfelet. Solna, Skatteverket 2006:7)

United States Census Bureau (2018). Selected Characteristics of the Native and Foreign-Born Populations 2012-2016. American Community Survey 5-Year Estimates. Retrieved 2018-01-29

van der Heijden P. G. M., van Gils G., Cruijff M. & Hessen D. (2006), Een schatting van het aantal in Nederland verblijvende illegale vreemdelingen in 2005, IOPS Universiteit Utrecht, Utrecht.

Weitoft G. R., Gullberg A., Hjern A. & Rosen M. (1999). Mortality statistics in immigrant research: method for adjusting underestimation of mortality. Int J Epidemiol, 28(4), 756-763.

Woodrow K. A. & Passel J. S. (1990). "Post-IRCA Undocumented Immigration to the United States: Assessment Based on the June 1988 CPS". In Bean, F. D., Edmonston, B., & Passel, J. S. (Eds.). (1990). Undocumented Migration to the United States: IRCA and the Experience of the 1980s (Vol. 7). The Urban Institute

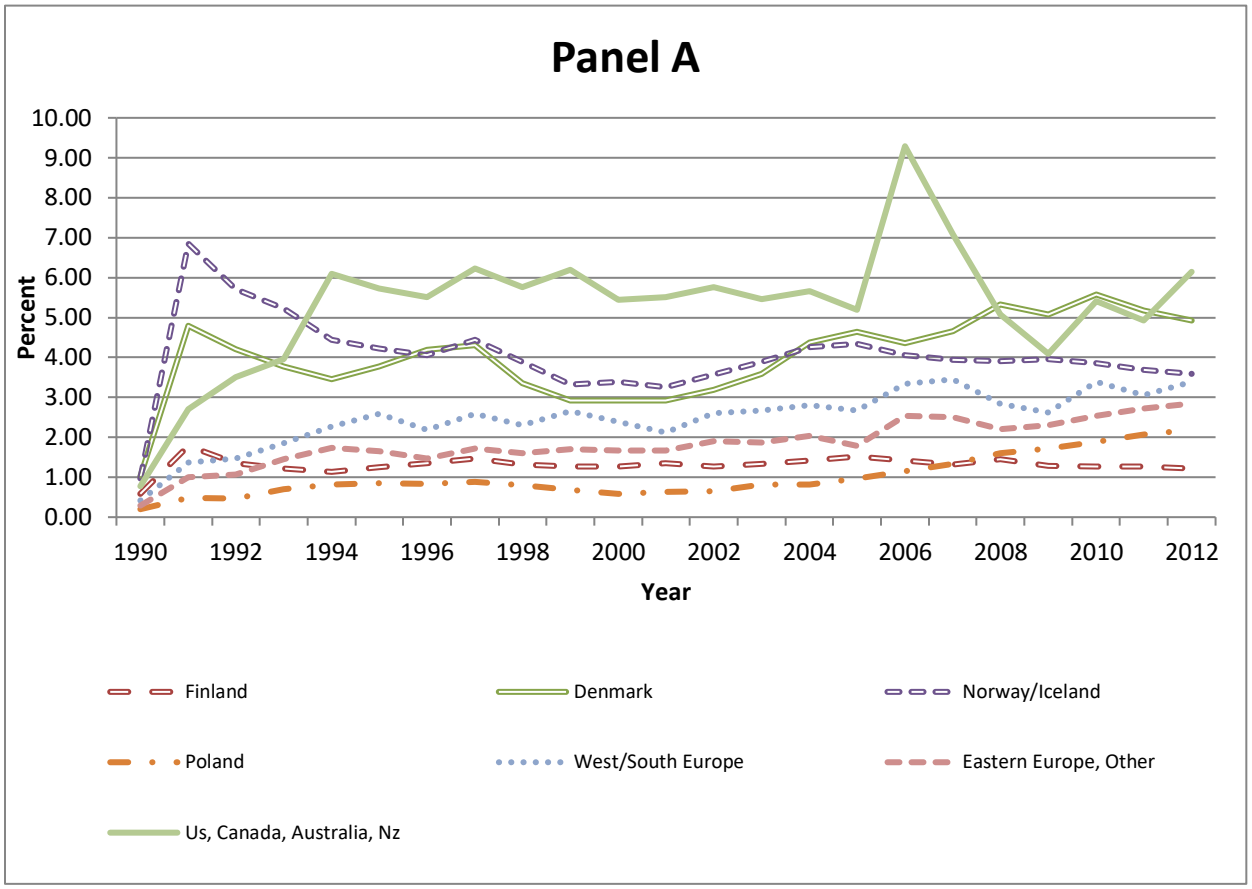
Appendix.

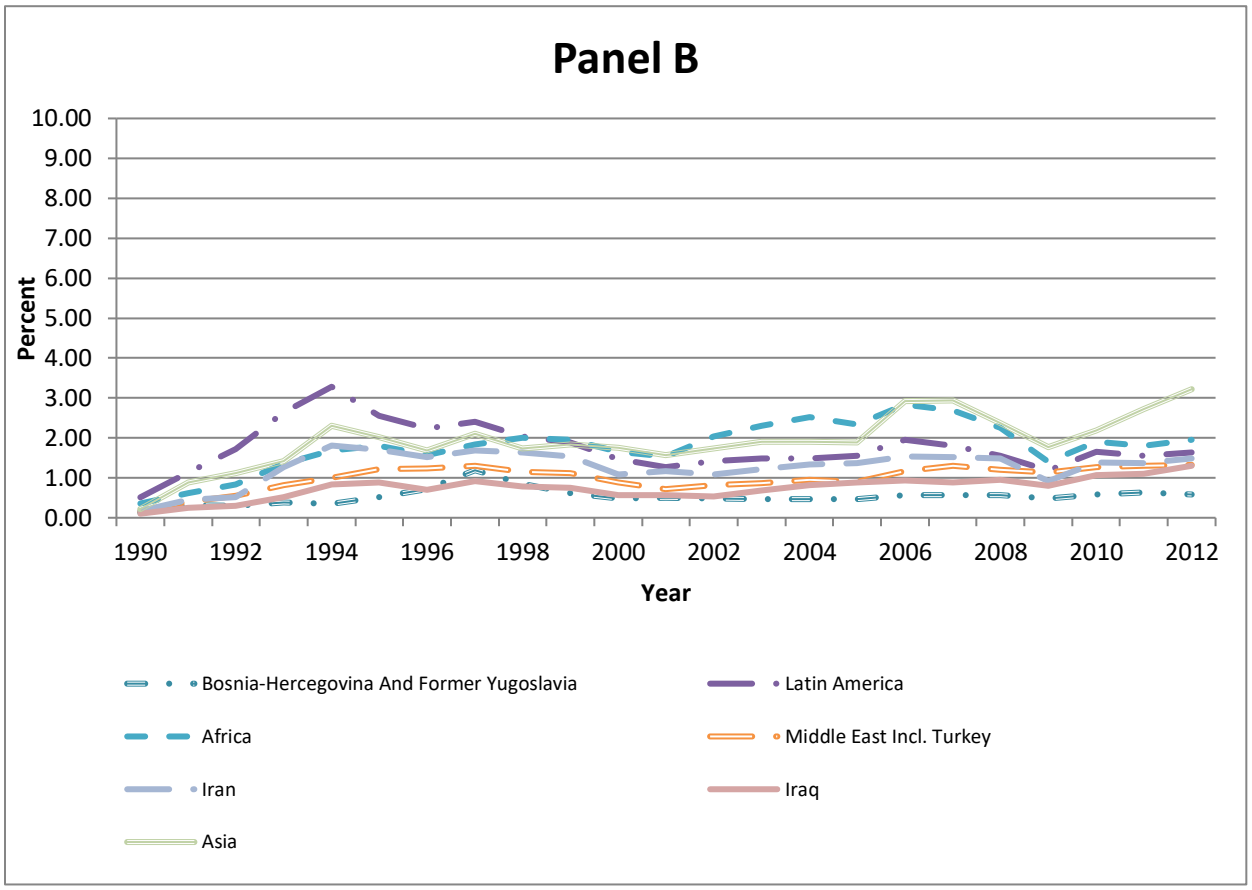
Table 1. The longitudinal *register-trace* approach.

	Indicators		Weight by SCB (2015:45)
	All indicators refer to year t, and only to those individuals not regarded active in the cross-sectional *register-trace* approach.		
1	Death the year after (t+1)	Indicates correct registration	2
2	Death, internal move, change in civil status or citizenship the year before and after the inactive year (t-1 and t+1)		2
3	Active again the year after and with no internal move within Sweden in between (t+1)		2
4	Foreign citizen who immigrated after age 60		2
5	Reason of residence permit: enough financial capital to support him/herself		2
6	Household (household) income over yearly national base amount (calculated in relation to consumer price index) (time t)		1
7	Emigration the year after (t+1)	Indicates over-coverage	3
8	Enrolled in tertiary education the year before (t-1), and a foreign citizen (time t)		2
9	Reason of residence permit: Studies		2
10	Immigration two years before (t-2), followed by a positive personal income the first year (t-1)		3
11	Positive personal income the year before (t-1), positive income the year after (t+1) and a new address (t+1)		2
12	Not any known address (time t)		2
13	Not registered in the Swedish Total Population Register the year after (t+1), without any notification on death nor emigration		3
14	A positive personal income the year before (t-1)		1
15	Reason of residence permit: Work		2
16	A registered death the year before (t-1)		3
17	A registered emigration the year before (t-1)		3
18	A registered immigration the year after (t+1)		3

All indicators correspond, but are not equal, to indicators listed by SCB (2015): 41-45. Weights are the same as the SCB weights.

Figure 6: Emigration rates by country of birth, years 1990–2012.





Source: Swedish register data. Calculations made by authors

Stockholm Research Reports in Demography
Stockholm University, 106 91 Stockholm, Sweden
www.su.se | info@su.se | ISSN 2002-617X

Stockholm University
Demography Unit