Working Papers
www.mmg.mpg.de/workingpapers

Max Planck Institute for the Study of Religious and Ethnic Diversity

Max-Planck-Institut zur Erforschung multireligiöser und multiethnischer Gesellschaften

MMG Working Paper 10-05 • ISSN 2192-2357

ALAN GAMLEN

International Migration Data and the Study of Super-Diversity

Alan Gamlen
International Migration Data and the Study of Super-Diversity

---

MMG Working Paper 10-05

Max-Planck-Institut zur Erforschung multireligiöser und multiethnischer Gesellschaften,
Max Planck Institute for the Study of Religious and Ethnic Diversity
Göttingen

ISSN 2192-2357 (MMG Working Papers Print)

Download: www.mmg.mpg.de/workingpapers

MPI zur Erforschung multireligiöser und multiethnischer Gesellschaften
MPI for the Study of Religious and Ethnic Diversity, Göttingen
Hermann-Föge-Weg 11, 37073 Göttingen, Germany
Tel.: +49 (551) 4956 - 0
Fax: +49 (551) 4956 - 170

www.mmg.mpg.de

info@mmg.mpg.de

Abstract

The purpose of this paper is to review ten prominent sources of data on international migration, specifically in light of their relevance to research on the 'superdiversification' of international migration in the post-World War II period, and in particular to the hypothesis that migration patterns involving large flows between few places have shifted to patterns involving smaller flows between more places. In addition to an introduction and conclusion the paper comprises two main sections. The first discusses the types of underlying source data from which global migration datasets are generally composed, highlighting their particular characteristics and the challenges of availability and compatibility which arise when combining them to create more comprehensive databases. The second section of the paper provides a brief review of several hundred words for each of ten major international migration datasets. The conclusion draws attention to three datasets of particular relevance to studying the superdiversification of migration: the OECD's SOPEMI Database, the UNPD's Flows to and from Selected Countries (2008 Revision), and the emerging World Bank-led Database of Global Bilateral Migration History. The paper ends by noting that while analyses of these databases can yield a macro-view of the diversification of international migration, micro-data are in the long run needed to probe the intricacies of superdiversity.

Author

ALAN GAMLEN is Postdoctoral Research Fellow at the Max Planck Institute for the Study of Religious and Ethnic Diversity (MMG), Department for Socio-Cultural Diversity, Göttingen, and ESRC Post Doctoral Fellow at the International Migration Institute (IMI), University of Oxford.

alan.gamlen@qeh.ox.ac.uk

Contents

Introduction

The purpose of this paper is to review ten prominent sources of data on international migration, in a digestible but still comparatively detailed format. This is part of a wider project to analyze and visually represent the 'superdiversification' of migration.

Steven Vertovec (2007) introduces the notion of 'super-diversity' 'to underline a level and kind of complexity surpassing anything … previously experienced'. According to Vertovec, conventional understandings of 'diversity' in Britain, Europe and elsewhere have focused on 'large, well-organized' groups from a few traditional places of origin, which tend to be treated in academic and policy discourses as relatively homogenous 'communities'. In large part owing to globalization, migration and related processes of social transformation, the apparent homogeneity of such groups is increasingly in question: existing cleavages in migrant and minority 'communities' have been compounded and cross-cut by new 'axes of differentiation' – including a kaleidoscope of new ethnic and national distinctions, and newer variables of interest such as legal status, religion, class, gender, age, and so on. Rather than conforming to a single primary identity, immigrant and ethnic minority populations are now bound to each other, to their 'hosts', and to distant compatriots by multiple strands of solidarity of varying thickness. This shift away from 'multiculturalism' based around discrete 'cultural communities', towards a more complex pattern of social ties and tensions – 'the diversification of diversity', as Vertovec describes it – is the focus of 'superdiversity' research.

Close to the heart of the superdiversity concept lies a hypothesis about migration: the hypothesis that, over the past few decades, patterns of international migration involving *many* migrants from and to *few* places have shifted to patterns involving *fewer* migrants from and to *more* places. The growing complexity and differentiation of global migration has received increasing attention in the migration literature in recent years (e.g. see Boyle et al 1998: 28; Castles & Miller 2003: 7-9), yet macro-analyses of these trends lack the support of truly comprehensive international migration data. This is by no means a fault of the analyses themselves, but wholly an upshot of the current lack of data that permit detailed examination of trends in the volume and direction of international migration flows in the postwar period.

The surprising lack of global migration data partly reflects the state of international cooperation over migration-related matters. Instead of a coherent global migration governance framework – one of whose functions would be to collect and

disseminate high quality international migration data – we have instead a 'fragmented tapestry' (Betts forthcoming) or a 'multilevel patchwork' (Gamlen forthcoming) of local, national and international organizations managing migration in disparate and overlapping ways. This situation is clearly reflected in the range of currently available migration datasets reviewed below. Rather than a coordinated international process leading to a comprehensive and authoritative source of global migration data, we have various local, national and international organizations collecting migration data using different and not always compatible techniques and procedures, resulting in a jumble of databases of varying quality. For the uninitiated, sifting through this untidy assortment can be a difficult and ultimately unrewarding task: there is a great deal out there, but much of it is of dubious quality, very little of it is organized systematically, and therefore there are many gaps and overlaps in coverage.

Against this background, the purpose of this paper is simply to review ten of the foremost sources of data on international migration in the postwar period. Aside from its main purpose of functioning as a guide to international migration data, particularly data relevant to the study of superdiversification, the paper also offers a fresh approach to integrating quantitative and qualitative methods in migration studies. Despite widespread acknowledgment of the need to study migration using mixed methods, the development of the mixed methods research toolkit is still very much underway, and ways of convincingly integrating methods are in short supply. Relatively few studies offer integrated approaches beyond rudimentary 'sequential' designs (Creswell 2009: 211-13) in which one method is first used to generate an overview, and the other is then used to 'drill down' into specific questions.

This study looks laterally at mixing methods: it uses qualitative methods to study the basic tool of the quantitative migration researcher – the dataset. Meetings and interviews were held with 29 expert collectors and users of international migration data based in the relevant global institutions, identified primarily by referral (see Appendix for full listing). These experts were asked, in semi-structured interviews, about the nature, uses and limitations of the datasets they relied on in their research. The meetings and interviews were conducted in November 2009-February 2010, by telephone and on fieldtrips to Geneva, New York and Washington DC. Additional experts provided advice via email. Key secondary and primary literature – particularly the documentation and metadata accompanying the datasets of interest – was also consulted and drawn on extensively below. It is hoped that this qualitative approach has been able to produce a kind of guide that both quantitative and qualitative migration researchers might find useful.

In addition to this introduction and a conclusion, this paper contains two sections. Section One discusses general issues around the compilation of migration source data. Section Two consists of individual reviews for ten leading databases recommended by experts. Each review contains a contextual description and a discussion of uses and limitations, in addition to specifying the type of data sources used, the historical and geographical extent of the dataset, which variables are included, and how to access the dataset. The conclusion briefly highlights which of the datasets under review seem best suited to studying the superdiversification of migration.

Section 1. Underlying Sources

Types of Sources

The *datasets* or *databases* covered in this report are compiled from a range of *sources*, including population registers, migration permit data, specific surveys (including household, labour force and border control surveys), and population censuses. Each of these has its own specific limitations, which will be briefly discussed in this section. Additional limitations are imposed when combining different sources into a single *database* / *dataset;* these are discussed in the next section.

First, however, it is useful to distinguish the scale at which data is collected. Data in which the unit of analysis is a geographical unit such as a province, country or region is referred to as *macrodata*, while data in which the unit of analysis is the individual or individual household is referred to as *microdata*. The central hypothesis of this paper calls for macrodata dealing with the volume and direction of migration flows between *countries*, but it is worth noting that studying aspects of superdiversity, with its underpinning concepts of multiple cross-cutting axes of differentiation among *individuals*, would generally seem best served by microdata.

Population Registers

Some countries require all residents to register with their local government, and the resulting data can be a rich source of information on the migrant population (UNDESA Statistics Division 1998: 25-27). The criteria for registration vary widely among countries, particularly with respect to the minimum duration of stay, presenting problems when comparing across countries (OECD 2009: 3). In some countries, dependants are not required to register and are therefore not captured in the data. Population registers also tend to undercount departures, as emigrants either over-

look or actively avoid deregistration. Like many international migration data sources, registers exclude undocumented migrants.

Permit Data

As the OECD notes (OECD 2009: 3), residence-permit applications and approvals constitute another important source of international migration data (also see UNDESA Statistics Division 1998: 28-32). There are many different kinds of permit, ranging from temporary work permits to permanent residence permits – not all of which are fully comparable across countries. Permit data have a number of important limitations. Such data are distorted when approved permits are not taken up by the applicant (leading to overcounts of migrants), or when backlogs of permit applications are cleared *en masse* even though migrants may have already entered the country (leading to a sudden spike in the time series). Permit data on duration of stay must be treated with caution: a one-year permit holder may not stay a full year, or may switch to another visa category and stay longer. Emigrating and returning citizens are excluded from permit data because they do not require permits. Like population registers, permit data by definition exclude undocumented populations.

Household, Labour Force, and Border Surveys

The characteristics and limitations of survey data are generally specific to the type of survey and the context of its implementation (UNDESA Statistics Division 1998: 32-35). Border surveys – which enumerate all entries and exits – are very accurate, but costly and exacting to implement. Consequently they tend to exist among developed 'settler' countries, and even there it is difficult to reliably match one countries exits to another's entries. (For example, the origin and destination information from border surveys can be complicated by stopovers, and intended durations of stay are not always reliable indicators of actual durations.) Household surveys by definition exclude those who have moved away – although attempts (of somewhat dubious quality) are sometimes made to gather information from households about their absent members. Surveys generally include some undocumented migrants. However, care should be taken when making inferences because the migrant sample size is almost always vanishingly small (OECD 2009: 6) and therefore highly subject to sampling error.

Censuses

Censuses usually either enumerate immigrants either by birthplace or nationality, and sometimes both – though it causes problems that definitions are not consistent

among countries (see below). Some censuses also give information on the respondent's previous place of residence and/or length of residence in the current region; when broken down by age and nationality, these measures can provide the tools to arrive at a definition of migrant. Census data are, of course, more comprehensive than other surveys, but they still systematically undercount migrants (who tend to overlook or avoid them) (Hugo 2006: 114), and they occur only infrequently – generally once a decade, precluding all but the most general historical study. Census sources generally only list the most important migrant source countries, and lump the rest into catchall categories like 'Other Africa', which are difficult to disaggregate retrospectively (Özden et al 2009).

Challenges to Combining Sources

In addition to source-specific limitations, international migration datasets compiled from a variety of sources are limited by issues of *availability* and *comparability*. Owing to these constraints, currently available global migration datasets are invariably composed from a patchwork of heterogenous sources which measure differently defined populations and are therefore neither comprehensive nor truly comparable.

Availability

The main challenges to data availability are geographical and historical gaps, and a restricted range of variables. Perhaps the two most obvious problems are that most countries do not seem to have collected migration data until relatively recently (with the result that even time series for traditional immigration countries like New Zealand typically only stretch back to the middle of the 20th century), and that most migration data comes from official sources which necessarily exclude undocumented migrants.

However, other important limitations also exist. There are major problems with the quality and quantity of international migration data from developing countries. In general the capacity of government statistical agencies in these countries is limited. Traumatic events such as conflict and environmental disaster, which affect developing areas more acutely, can create gaps in national time series data. Much of the migration data from developing countries is thus of dubious quality if not altogether missing.

Shifting borders are another source of major gaps in international migration time series (Özden et al 2009). In addition to driving changes in the actual volume, direc-

tion and composition of these flows, border changes abruptly redefine internal mobility as international migration, making it difficult to backdate flows based on current borders. The border changes that accompanied decolonization generally occurred before the start of most international migration time series, but those accompanying the end of the Cold War, particularly the disintegration of the USSR and Yugoslavia, present a major challenge to the study of postwar global migration.

Where data distinguishing migrants from non-migrants is available, it is often only in a highly aggregated form. As mentioned, censuses often treat entire continents as single source regions. Even countries with comparably developed migration data – such as the USA – do not always enumerate outflows. Disaggregation by demographic, labour market and other characteristics is rare, and there are always trade-offs between comprehensiveness and detail. For example, on one hand the UNPD's Global Migration Database,[1] one of the most comprehensive sources of migration data currently available, struggles to achieve disaggregation by birthplace and citizenship, let alone age and sex. On the other hand, the IPUMS International[2] database of census samples contains extremely detailed microdata, but is restricted to 44 countries and contains only samples, which do not allow tracking of aggregate trends.

A cumulative result of these difficulties is to make migration datasets look patchier the further back in time they go. This makes studying migration in developed settlement countries much easier than anywhere else.

Comparability

Lack of comparability among different national migration data is a perennial problem for the study of international migration. In 1872 the International Statistical Institute recommended the harmonization of international migration data (Boyle et al 1998: 39), and various international agencies have made periodic recommendations reaffirming and elaborating on this recommendation ever since – for example, the UN has made recommendations every decade or so since World War II (Bilsborrow et al 1997; also see Simmons 1987; UNDESA Statistics Division 1998).

Many of the interviewees and sources consulted in this review highlighted that, since the last set of UN recommendations in 1998, there has been a significant improvement in the quality of international migration data. The need for more comparable international migration statistics has been a central topic of discussion at

1 See http://esa.un.org/unmigration/

2 See https://international.ipums.org/international/

various key multilateral meetings and forums such as the UN's annual Coordination Meetings on International Migration (of which there have now been eight), the Global Forum on Migration and Development, and the discussions of the Global Migration Group.[3] Numerous NGOs and think tanks have also begun to publish guides to international migration – which include a set of five recommendations by a MacArthur-funded high-level Commission on International Migration Data for Development Research and Policy published by the Center for Global Development in 2009 (Santo Tomas et al 2009), and a pocket guide to migration sources by the Migration Policy Institute (Batalova et al 2008). These are only a few of the relevant initiatives that have either taken place or are currently underway to improve migration data.

However, owing to different national capacities and priorities, as well as lack of coordination among international agencies, this goal is still far from realization. The most important challenge to comparability is lack of agreement over the basic definition of 'migrant'. In some countries migrants are defined as foreign-*born* people, whereas in other countries they are defined as foreign *citizens* (Parsons et al 2007). The former definition includes only first generation migrants, whereas the latter may include their children and in some cases even their grandchildren. Birthplace is usually preferred by migration researchers, because it is a stable characteristic and one that unequivocally proves that movement has taken place.

In addition, there is no clear consensus on how to define different categories of migrant. Although the UN has suggested that stays of up to three months should be classified as tourism, up to one year as short-term, and over one year as long-term (UNDESA Statistics Division 1998: 18), only a few countries have complied and in practice definitions differ widely among countries. Many Arab states, for example, insist they have 'temporary workers' rather than 'migrants'. Similarly, different countries use different definitions of asylum seekers, workers, dependants and students, hindering cross-country comparison. For example, some countries count asylum seekers at the time of application approval, which may occur some time after the applicant entered the country. Some countries count only primary applicants and not the dependants they may bring with them (OECD 2009: 5).[4] As a result of discrepancies in definitions, migration databases are often forced to elide different populations under the catchall heading of 'migrant.'

---

3 See http://www.un.org/esa/population/migration/

4 Also see http://www.unhcr.org/pages/4a013eb06.html

Section 2. Ten Existing Databases

Notwithstanding the limitations of specific source data, a wide range of organizations have compiled databases relating to international migration. At the most basic level these contain information on national flows and stocks of migrants aggregated to the international level. In a number of cases, disaggregation by various demographic characteristics is also included, and in a few cases further data such as labour market information is also available.

This section reviews the databases and outlets recommended by the expert interviewees. Ten key databases are reviewed in some detail, and a further five outlets which compile third-party data are reviewed in brief. In addition to a contextual description and a discussion of uses and limitations, each review contains details of the type of sources from which the database is compiled, the historical and geographical extent of the database, the variables included, and accessibility issues. The heading of each review is a hyperlink to the data and / or accompanying documentation (press control + click to leap directly to the associated webpage).

The reviews are arranged in a sequence which highlights both their relevance to the superdiversification hypothesis and their inter-relationships. So, for example, stock data is reviewed first because, as shown below, it is currently more widely available than flow data and therefore most useful for analyzing the proliferation of small migration corridors posited by our main hypothesis. These stock databases are then reviewed in historical sequence, to highlight how each effort has built on the lessons of those compiled previously. If this arrangement seems rather ad hoc, it is: a logical order might have presented itself if the datasets themselves were logically coordinated – but, as mentioned, this is not the case.

SUSSEX GLOBAL MIGRANT ORIGIN DATABASE

Type:	census stocks with supplements
Historical coverage:	cross-sectional c2000
Geographical coverage:	global (226 countries)
Variables:	birthplace, nationality, destination
Accessibility:	publicly available online

Background and description:

This project began around 2003, coordinated by staff at the Migration and Development Research Centre at Sussex University (Parsons et al 2007). The core of this database is a 226x226 cell matrix of global bilateral migrant stock circa 2000, where

each cell represents the bilateral migrant stock between two countries (i.e. Australians in New Zealand, Ghanaians in the UK, and so on). The data has been compiled from all available national censuses in the '2000 round' (i.e. 1996-2004). Raw data is only available for 162 of the 226 countries; estimates are used sometimes to fill the remaining 64 countries.

Several versions of the matrix exist. Each edition consists of two 226x226 matrices: one based on the 'birthplace' definition of migrant, and another based on the 'nationality' definition. There are four successive 'editions', which become increasingly complete but inaccurate, as estimation techniques are progressively used to fill gaps.

Uses and limitations:

The Sussex matrix provides a one-time cross-sectional 'snapshot' of migration around the year 2000. There is no historical component, nor any further information about the characteristics of migrants, but this represents a first attempt to compile a comprehensive picture of bilateral migration corridors using stock data from national censuses – an approach which, as seen below, has since been quite fruitful.

Even high quality censuses present various problems for studying migration. They systematically undercount migrants, who may wish to avoid state surveillance or feel separate from the population being surveyed. Because census rounds are staggered (the 2000 round stretches from 1995 to 2004), data is not all from the same year. They count stocks, not flows, and therefore date of arrival is unknown: those who arrived one year ago are not distinguished from those who arrived 40 years ago, as Graeme Hugo points out (interview).

Not all censuses are high quality, if they exist at all. There are geographical gaps, particularly for developing countries. Many censuses do not enumerate every country of origin individually; instead they list only the most common and combine the remainder into regional catchall categories such as 'Other Africa', which are disaggregated using estimation techniques – all of which are open to substantive criticism – in later editions of the Sussex matrix.

World Bank Bilateral Migration Matrix

Link to accompanying paper (press control + click). Note: the construction of the database is described in Appendix A from page 37.

Type: census stocks with supplements
Historical coverage: cross-sectional c2000

Geographical coverage: global (212 countries)
Variables: origin (birthplace and nationality undistinguished), destination
Accessibility: publicly available online

Background and description:
This 212x212 cell matrix revises the Sussex Global Migrant Origin Database in order to facilitate econometric modeling (Ratha & Shaw 2007). Migrant stock data for more than 50 countries is updated or replaced, and attempts are made to 'improve' the Sussex database by scaling aggregated stocks culled from individual national senses to UNPD stock figures, and introducing the concepts of 'unidentified migrant' (whose origin and destination are known), and 'combined migrant stock' (where birthplace and nationality are treated as equivalent).

Uses and limitations:
This dataset has fewer countries than Sussex, but is more useful for econometric modeling. Whereas the Sussex database was more useful for studying small island states, the World Bank matrix is stronger on Latin American data. However, the World Bank matrix is significantly less accurate than Sussex and therefore only appropriate for macro-level analysis. Firstly, the concept of a 'combined migrant stock', which conflates birthplace and nationality, is problematic: the former refers to a fixed characteristic of the first generation whereas the latter is a highly fluid characteristic of the first, second and even third generations. Secondly, 'unidentified' migrants (i.e. those whose origins were unknown due to aggregation in groupings such as 'Other Africa' (see above) were simply allocated to the catchall categories: 'Other North' and 'Other South'. This suits the specific purpose of the dataset – for studying South-South migration and remittances – but is otherwise problematic.

World Bank Database of Global Bilateral Migration History

Type: census stocks with supplements
Historical coverage: c1960-c2000 at 10-year intervals
Geographical coverage: global (226 countries)
Variables: birthplace, nationality, destination (in progress: sex, age)
Accessibility: in progress, not publicly available. In talks with World Bank team.

Background and description:

This bilateral database is still under construction (Özden et al 2009), and the name given here is not official but a descriptive label I have applied for the sake of convenience. In essence the database aims to replicate the Sussex matrix for every census round since 1960 (i.e. five rounds in total), thereby achieving broad historical treatment in addition to comprehensive geographical coverage. It draws on a large and highly diverse range of sources, particularly census records culled from libraries and archives around the world, and the United Nations Population Division's huge (but still slightly unwieldy) Global Migration Database (see below). Data becomes patchier further back in time, particularly for developing countries, but updated estimation techniques are being used to interpolate missing data, whilst taking care to clearly document any transformations of raw data. Attempts are also currently underway to disaggregate the data by gender and age. The dataset contains around 500,000 individual cells of data. Some 40% of the cells are filled with raw country-level data, while around 20% are filled with aggregate raw data (which needs disaggregating), and the remaining 40% of the data is still missing.

Uses and limitations:

When complete, this dataset will provide a blurry, stop-motion-like but nevertheless animated and unprecedentedly comprehensive picture of international migration in the post World War II period. It will trace the development of every bilateral migration 'corridor' in the world, at roughly ten-yearly intervals.

At its current stage of construction, the database still contains some significant gaps. Firstly, shifting borders after the Cold War transformed internal migration into international migration, and therefore studying the evolution of flows across present-day borders thus requires data on internal movement within the USSR and Yugoslavia. In the former case, this is particularly problematic not only because raw Soviet census data can be difficult to access, but also because most Soviet censuses did not include a birthplace question.[5] The absence of raw birthplace data for major receiving countries such as France, Germany and Italy constitutes a second major gap

5 A birthplace question was included in the draft of the questionnaire for the infamous suppressed 1937 Soviet Census (see, for example, http://en.wikipedia.org/wiki/Soviet_Census_%281937%29), the organizing officials of which were promptly executed for exposing sensitive information on famine mortality in the period (Merridale, C. 1996. The 1937 Census and the Limits of Stalinist Rule. *Historical Journal* 39: 225-40). However, Stalin himself, who was allowed to edit the final questionnaire, removed this (and many other questions) and it was not reinstated until 1989.

– one that may in future be addressed through estimation techniques currently being developed at the Max Planck Institute for the Study of Religious and Ethnic Diversity. A third major gap is the absence of data for the Gulf Cooperation countries.

As in the Sussex matrix, one of the most important challenges to compiling the database is disaggregating stock figures in 'rest of' categories such as 'Other Africa' (see above). A specific process has been developed for this purpose. Firstly, the rest-of category is identified as one of several types, based on what countries it includes. Next, attempts are made to locate data disaggregated by origin country pertaining to other years in the same receiving country; if such data are available, they are used as benchmarks to estimate the composition of the 'rest of' category in the missing year.[6] If data for another year is unavailable, proportions from the summation of all years are used. If this is not possible, proportions from inflow data at the regional level (e.g. whole of Europe) are used where available. Entirely missing census years are interpolated by scaling the data to the UN stock totals, assuming linear growth trends in previous bilateral stocks based on UN growth rates.

The dataset will inevitably suffer from many of the same weaknesses as its forerunners, such as different dates, undercounting, and data 'smudging' through estimation techniques (even if these are improving significantly, for example by discarding the controversial 'entropy' measures for estimating birthplace from nationality). Moreover, some of these problems will be compounded for earlier historical periods. However, there will be rigorous documentation clarifying exactly how the data was created, to ensure it does not become a 'black box'.

UNPD Trends in International Migrant Stock, 2008 Revision

Type:	stocks from censuses, population registers and surveys
Historical coverage:	1960-2010
Geographical coverage:	global (c221 countries)
Variables:	destination only
Accessibility:	publicly available online and in CD-Rom

6 For example, say there is only a 'rest of Africa' category for the 1980 UK census, but it is also known that ten years earlier in 1970, 80% of all UK immigrants from countries comprising this composite category were actually from Zimbabwe. It is then assumed that Zimbabweans also comprised 80% of the UK's 'rest of Africa' category in 1980.

Background and description:

This is a regularly updated set of tables enumerating the number of migrants in each of the world's countries since 1960 (half the tables only go back to 1990) (UNDESA Population Division 2009b). Disaggregation by age and sex is available in some instances, as detailed in the documentation. Wherever possible the 'birthplace' definition of migrant is used; this was possible in 179 countries, but not possible in 42 countries, where the nationality definition was used instead. The documentation accompanying the dataset details which definition was used in each case. The dataset is revised every five years or so; 2008 is the most recent revision – as Bela Hovy explained (interview), it draws on the new Global Migration Database and therefore contains 40% more sources than the 2006 revision (2006 used around 700-750 stock sources whereas the 2008 version uses some 1,200).

Uses and limitations:

This data is easily accessible and regularly updated, and is useful for tracking macro-level trends. These can then be used as benchmarks for more detailed analysis using more detailed data. For example, as Ronald Skeldon explained (interview), at various points during the compilation of the Sussex and World Bank matrices, the Trends in International Migrant Stock were used to calibrate migrant counts aggregated from a diverse range of sources: if estimates from a diverse range of sources for migrants from all origin countries living in, say, the UK tallied up to around the same number as the Trends gave for the UK's total migrant stock, researchers knew that the heterogeneity of their sources was not majorly skewing their estimates, and any large discrepancies could be identified for closer scrutiny.

However, beyond such broad, macro-level analysis, the analytical potential of the stocks is very limited. The central limitation is that data is not disaggregated by country of origin: the data can tell us that there were X number of migrants in Country X in year X, but not where any of these migrants originated. This limitation makes it impossible to distinguish the size and direction of migration flows, which is the basic requirement for studying the proliferation of migrant-source countries in recent decades.

UNPD Flows to and from Selected Countries, 2008 Revision

Type:	flows, from population registers and border statistics
Historical coverage:	1970-2010 for some countries; becomes comprehensive around 1990
Geographical coverage:	29 industrialized countries
Variables:	birthplace, nationality, destination, duration of stay
Accessibility:	publicly available in CD Rom format

Background and description:

This dataset provides annual inflows, outflows and net migration for 29 key migrant receiving countries, stretching as far back as 1970 (for nine classical immigration countries) (UNDESA Population Division 2009a). Seven mainly Northern and Southern European countries have data to 1980, while 11 mainly Eastern European countries only have data since c1990, and two only have data starting in the current decade. Data for most countries is based on population registers and, in a few cases, border statistics. Various definitions of migrant are used (birthplace, nationality, etc) but these are clearly documented in each case. The previous revision of this dataset took place in 2005 and only includes 16 countries, but the consolidation of the Global Migration Database (see below) in recent years made it possible to augment and refine these sources considerably.

Uses and limitations:

Because it includes annual rather than decennial time series data, and includes duration of stay, this dataset provides a significantly more detailed picture of migration than, for example, the World Bank Database of Global Bilateral Migration History. However, what it gains in detail it loses in historical and geographical coverage: only 29 countries are covered, and the database only becomes comprehensive around 1990.

The 2008 Revision is more standardized than the 2005 Revision, making it somewhat easier to compare countries. However, the problem of defining 'migrant' is knottier for flow data than for stock data: in addition to the question of birthplace vs. nationality, there is the question of duration between entry and exit. Only 11 countries follow the UN recommendations to define one year as the cut-off between short-term and long-term migration; the remaining countries use widely varying definitions for different categories of migrant (foreigner, citizen, immigrant, emigrant and so on) (UNDESA Population Division 2009b: 1-3). Moreover, durations themselves are often based on policy categories or migrant intentions, rather than migrant beha-

viour. In the case of permit data, for example, a one-year permit holder may not stay a full year, or may switch to another visa category and stay longer. In addition, some data is based on residence permits; because returning and emigrating citizens do not require these they are excluded from figures, and the clearance of backlogs in permit approvals may cause the appearance of a spike in flows that were in fact steady. A significant number of people choose not to migrate even though they have been granted permission. As mentioned in Section One, population registers – the main source of data – tend to undercount departures, and criteria for registration itself varies widely among countries.

UNPD Global Migration Database

Type:	Stocks only, from a diverse range of sources
Historical coverage:	varies widely
Geographical coverage:	global
Variables:	depends on country, but typically: birthplace, nationality, destination; occasionally has sex and age
Accessibility:	accessible online to 'key partners' of UNPD by registration

Background and description:

This is perhaps the most comprehensive compilation of international migration data sources currently available (see http://esa.un.org/unmigration/). As Bela Hovy explained (interview), this database has been created by combining the massive migration stock database of the UN Department of Economic and Social Affairs (which is constantly checked and updated) with the UN Population Division's own files accumulated over many years but often existing only in hard copy versions in staff filing cabinets. Most of the data has come from individual governments, and thus follows a wide variety of local definitions that sometimes change over time. The sources have been digitized and posted online, with accompanying descriptions of the variables contained in each series. The website notes that, because it is still in a testing phase, the resulting database is available to 'key partners' of UNPD, through registration.

Uses and limitations:

The major limitation to the dataset is its inconsistency in terms of geographical and historical coverage and comparability between regions and periods. The data are from heterogenous sources and are stored in unharmonized form: although they are

documented clearly in each case, definitions and historical/geographical coverage vary considerably among countries, leaving many gaps. Because spreadsheets must be downloaded separately, it is not very easy to compile the data for aggregate analysis. Owing to its early stage of development and limited accessibility, this dataset has not been extensively explored and its main uses are perhaps yet to be discovered, but one of its most valuable functions maybe as a kind of 'reference text' database. It is certainly an unparalleled resource for researchers constructing databases for particular purposes (as it has been for the World Bank-led team currently compiling the Database of Global Bilateral Migration History), or as a starting point for in-depth studies of particular regions or periods – much as one might consult encyclopedia entries at an early stage in researching a particular topic.

OECD SOPEMI Database

Type:	flows, stocks and naturalizations from population registers, work permits, specific surveys (including household surveys and border statistics)
Historical coverage:	becomes comprehensive around 1990
Geographical coverage:	basically OECD
Variables:	inflows, outflows, birth country, citizenship, acquisition of nationality, asylum seeker entries, labour market data
Accessibility:	publicly available online, more data available by subscription to OECD.Stat

Background and description:

SOPEMI is an acronym for the French title meaning Continuous Observatory on Migration. The Observatory is empowered to recommend but not enforce standards in data collection and harmonization. The data for each country are provided by a designated 'reporter' who fills in the OECD's questionnaire. Local data does not always fit the standard questionnaire: in practice, therefore, the character of data is driven by local factors in each member state, and this places limitations on harmonization (OECD 2009).

The SOPEMI database contains tables of annual stocks; inflows, outflows and net migration; and naturalizations for OECD countries, stretching back as far as 1984 (for around a third of member states) – although the database only becomes really comprehensive around 1990. Data come from population registers, residence and

work permits, censuses and specific surveys (such as the labour force and household surveys or border statistics).

Tables in the 'A' series only contain aggregate in and outflows, whereas 'B' series tables are disaggregated by nationality, revealing the origin / direction of flows. Only OECD destinations are included, but wherever possible the top 15 origin countries are listed whether or not they are OECD members – although this unfortunately leaves a large residual 'Other' category which cannot be further disaggregated (OECD 2009: 2). Additional tables are available to subscribers on the OECD.Stat website; this often provides data back to 1975; in some individual cases, for example the USA, these contain very long time series.

Uses and limitations:

This is one of the most widely used and reliable sources of international migration data in existence; it has been continuously revised, refined and extended over the past two decades and is constantly improving. It constitutes a kind of industry standard, and there are various efforts to create similar observatories around the world.

However, SOPEMI has both geographical and historical limitations: it is restricted to OECD destinations, disaggregated by the top 15 source countries. This makes it oblivious to large South-South flows, and unable to reveal all but the largest migration corridors. Moreover, it only becomes comprehensive after 1990, limiting scope for historical research. The data is presented in more or less standardized tables, but because it derives from individual country reporters using local definitions, the same problems of comparability as mentioned previously still apply. As in most other cases, illegal migration is not reported.

The limitations of flow data in general have been discussed previously, but are worth reiterating. SOPEMI Stock data derives from population registers, residence permits, labour force surveys and censuses. Population register stocks are inflated by emigrants who failed to deregister before departing, but deflated by undercounts of dependants who enter on a parent's or partner's permit. Census data are relatively comprehensive – even to the extent of counting parts of the undocumented population – but still systematically undercount migrants and occur too infrequently to allow detailed historical analysis. Although they also include some undocumented migrants, specific surveys such as the household survey can be problematic because the migrant sample size is almost always very small (OECD 2009: 6).

OECD Database on Immigrants in OECD Countries (DIOC and DIOC Extended)

Link to DIOC – Extended data and accompanying documentation (press control + click).

Type:	stocks from censuses, population registers and labour force surveys
Historical coverage:	cross sectional, c2000
Geographical coverage:	DIOC includes OECD destinations, over 200 countries of origin; DIOC Extended includes 60 destination countries (the extra countries are largely in Latin America, with four extra in Africa and two in Europe)
Variables:	age, gender, duration of stay, labour market status, occupation, industry sector, field of study, educational attainment, birthplace
Accessibility:	publicly available online, more data available by subscription to OECD.Stat

Background and description:

The DIOC database is similar to the Sussex matrix – it is a 'snapshot' of more than 200 countries around the year 2000 – but it contains fewer countries and more variables, particularly concerning the demographic and labour market characteristics of migrants (see above) (OECD 2008).

Uses and limitations:

Like other OECD databases, it is compiled from censuses and population registers supplemented by labour force surveys, with their attendant limitations. Because of its detail on educational attainment, the database is particularly useful for studying issues related to highly skilled migration and 'brain drain'.

United Nations High Commission for Refugees (UNHCR) Statistical Online Database

Type:	UNHCR country reports, based on registrations
Historical coverage:	1971 to present
Geographical coverage:	global (24 industrialized countries provide estimates only)
Variables:	refugees, asylum-seekers, returned refugees, internally displaced persons, returned IDPs, stateless persons, others of concern to UNHCR
Accessibility:	publicly available online

Background and description:

This database tracks trends in the population of concern to UNHCR, which includes refugees and returned refugees, asylum-seekers, internally displaced persons (IDPs) receiving UNHCR assistance, returned IDPs, stateless persons, and 'others of concern' to UNHCR.[7] The data is reported by country offices of UNHCR, and is constantly being updated and improved – as are data collection procedures. Data prior to 2007 include refugees who entered as part of a resettlement programme. The UNHCR mandate does not cover Palestinians; this data must be obtained from the United Nations Relief and Works Agency for Palestinian Refugees in the Near East (UNRWA).

Uses and limitations:

Data generally come from administrative sources in the host country and are therefore shaped by similar factors (e.g. see OECD 2009: 5). Several interviewees pointed out that, in many ways, refugee data is now much better than other migration data: it is collected every year, for each country, includes stocks, flows, characteristics, location in the host country.

There are both coverage and comparability issues with UNHCR data. Firstly, most data comes from registrations of people requiring UNHCR assistance, and the figures therefore suffer from undercounts: they generally derive from counts made in refugee camps, where tallying is relatively easy; 'self-settled' refugees in urban areas are undercounted because they are generally more integrated and require less assistance. For related reasons, industrialized countries do not generally count refugees separately from other migrants, leaving no reliable estimation mechanism. Secondly, although the 1951 Convention is very clear about the legal definition of a refugee, definitions are not entirely standardized over time. It took some time to impose reliable data collection on the strict definition itself, and now a new category for 'refugee-like situations' has been added. There is considerable debate over the limits of the population of concern to UNHCR, and these definitions differ widely among countries.

As with other permit data, family members of applicants are often excluded, and the recorded time of permit approval is often different from the unrecorded actual date of entry (OECD 2009: 5). Anomalies may arise from dual citizenship: for example, as Liliana Carvajal pointed out (interview), the USA appears to be a major source of refugees in terms of nationality, but in fact this reflects the fact that these

7 See http://www.unhcr.org/pages/4a013eb06.html

people have naturalized as US citizens. In addition, aggregate figures do not always agree with disaggregation by country of origin, because the later include appeals as well as initial applications, and it is not straightforward to separate the two.

Internally Displaced People are not discussed in this report, although it may be useful to note that the Internal Displacement Monitoring Centre, funded by the Norwegian Government and based in Geneva, is generally acknowledged as a leading source on internally displaced people. They produce a monthly online publication, and, as Stephen Castles noted (interview), their figures are often very different from UNHCR's. Link to Internally Displaced Persons Database and accompanying documentation (press control + click).

International Public Use Microdata Statistics (IPUMS)

Type:	samples from 130 censuses around the world
Historical coverage:	1960 to present
Geographical coverage:	sample of 279 million people across 44 countries, including non-OECD
Variables:	microdata; varies by year and country, but includes fertility, nuptiality, life-course transitions, migration, labour-force participation, occupational structure, education, ethnicity, and household composition
Accessibility:	publicly available online

Background and description:

IPUMS-International collects and harmonizes 'microdata' (data on individual persons and households) from around the world. This is achieved by taking samples from local census data, which are then 'harmonized' (coded and documented consistently across countries and over time).[8] It is already the world's largest database of publicly accessible census samples, and more samples are being acquired every year.

Uses and limitations:

IPUMS provides an unparalleled database of microdata in which variables have been harmonized across countries (including, unprecedentedly, non-OECD countries), allowing comparative study of characteristics of individual migrants around the world. Whereas census data is usually available only in aggregated format, microdata provides the information on individuals and individual households that was

8 See https://international.ipums.org/international/

originally collected on individual census forms, allowing researchers to build their own tables for specific purposes.

The main limitations to IPUMS migration data are that not all countries currently participate, and that the data are only samples from censuses and therefore do not allow examination of aggregate trends. IPUMS do not 'clean' the data they receive from censuses, and therefore they do not correct for flaws inherent in country reporting – although they do document the reporting format thoroughly.

Conclusions: Tracing the Diversification of Diversity

Having briefly reviewed ten leading databases on postwar global migration, it is now time to draw some conclusions about which of them are likely to be of most use for the specific analytical goals of studying superdiversification, and in particular for examining the hypothesis that 'migration has shifted from patterns involving large numbers of migrants to and from a few places, to patterns involving fewer migrants to and from a larger number of places'. This hypothesis calls for macro (i.e. country-level) data, although it is certainly worth bearing in mind that the broader notion of superdiversification along multiple axes of difference suggests a research focus that cannot be captured by conventional 'macro' categories, and which requires an understanding of the characteristics of individuals. This kind of study requires microdata at the level of the individual or the individual household. Bearing all this in mind, four datasets present themselves as particularly salient in studying the superdiversification of migration: the SOPEMI database; the UNPD database of Flows to and from Selected Countries (2008 Revision); the new World Bank Database of Global Bilateral Migration History; and the IPUMS International database.

The SOPEMI database contains both stocks and flows, is detailed, accessible, constantly updated and improved, well respected, and widely used in all kinds of serious research on global migration, and in this sense it should probably be consulted at the starting point of superdiversification research. However, SOPEMI is geographically restricted to OECD destinations and their top 15 migrant source countries, making it impossible to thoroughly assess the superdiversification hypothesis at the global level: the hypothesis calls for analysis of the proliferation of numerically smaller migration corridors, but in the SOPEMI data, only the 15 largest corridors are legible; the smaller corridors are amalgamated in a catchall category for the rest

of the world. Moreover, as SOPEMI only becomes comprehensive around 1990, it is difficult to examine long-term historical changes in patterns of diversity. Therefore, whilst SOPEMI is undoubtedly a key benchmark against which to measure findings, superdiversification research also needs to draw on additional sources.

The UNPD's database of Flows to and from Specific Countries overlaps somewhat with SOPEMI, as it covers 29 mainly developed countries. Rather than relying on the SOPEMI reporting system, UNPD draws on its own large databases and those of UNDESA, compiled from a very wide range of sources, often with longer time series and more detail than SOPEMI flows data – although it is more restricted insofar as it contains only flows and not stocks. Like SOPEMI, this database is constantly updated and improved, and because this takes place under United Nations auspices, the prospect of an eventual expansion in geographical coverage seems promising.

One of the most exciting datasets for analysis of our superdiversification hypothesis is the World Bank's emerging Database of Global Bilateral Migration History. Because it relies heavily on census data, it reveals stocks rather than flows, lumping together new arrivals with long-term settlers and even native-born people with migrant ancestry. Moreover, unlike the databases just mentioned, it only provides data at roughly ten-yearly intervals, providing at best a freeze-frame-like historical picture of migration. However, what it lacks in detail, this database makes up for in comprehensiveness: when it becomes publicly available it will paint, in broad brushstrokes, the evolution of every bilateral international migration corridor in the world since 1960. This type of data is well-suited to examining the hypothesized proliferation of smaller, more diverse migration corridors.

Finally, although the main hypothesis of interest in this paper calls for country-level data, the notion of superdiversification along multiple axes of difference suggests a focus that cannot be captured by data that homogenizes migrants into conventional national groups, but which instead requires an understanding of the characteristics of individuals. This kind of study requires microdata, such as that found in the IPUMS International database. Once the broad patterns of superdiversification have been mapped out using macrodata, the future of superdiversity research seems to lie in this type of detailed, individual-level information.

Despite numerous recommendations made over more than a century, and notwithstanding substantial recent improvements, it is often pointed out that we know more about textiles and cell-phones crossing borders than about people crossing borders. This is the somewhat disappointing reality of current international migration data. This situation seems unlikely to change without wider changes in the way inter-

national migration is governed at the global level. Greater institutional harmonization among states and international organizations tasked with migration matters seems a prerequisite for greater harmonization in the collection of global migration data. By bringing together information on ten of the leading examples of international migration data, this brief review modestly chips in towards that broader goal.

Acknowledgements

Sincere thanks are due to all the experts listed in the Appendices, who generously shared their time and knowledge. I am also grateful to the Max Planck Institute for the Study of Religious and Ethnic Diversity in Göttingen for funding this study, and in particular to Steven Vertovec and Norbert Winnige who are partners in the project of which it forms a part. I also acknowledge the International Migration Institute at the University of Oxford, and particularly Robin Cohen, for providing a stimulating working environment. Finally, I am grateful to Dorzhi Dondukov, Alexey Bessudnov and Alisa Voznaya for their assistance locating and translating data sources.

References

Batalova J, Mittelstadt M, Mather M, Lee M. 2008. *Immigration: data matters*. Washington DC: Population Reference Bureau / Migration Policy Institute.

Betts A. Forthcoming. Introduction. In *Global Migration Governance*, ed. A Betts. Oxford: Oxford University Press.

Bilsborrow RE, Hugo G, Oberai AS, Zlotnik H. 1997. *International Migration Statistics: Guidelines for Improving Data Collection Systems*. Geneva: International Labour Office.

Boyle P, Halfacree K, Robinson V. 1998. *Exploring Contemporary Migration*: Longman.

Castles S, Miller MJ. 2003. *The age of migration*. Basingstoke: Palgrave Macmillan.

Creswell JW. 2009. *Research design : qualitative, quantitative, and mixed method approaches*. Los Angeles; London: SAGE.

Gamlen A. Forthcoming. Diasporas and emigration states in the global governance of migration. In *Global Migration Governance*, ed. A Betts. Oxford: Oxford University Press.

Hugo G. 2006. An Australian Diaspora? *International Migration* 44: 105-33.

Merridale C. 1996. The 1937 Census and the Limits of Stalinist Rule. *Historical Journal* 39: 225-40.

OECD. 2008. *A Profile of Immigrant Populations in the 21st Century: Data from OECD Countries*: OECD.

OECD. 2009. International Migration Data 2009: Statistical Annex. OECD. Available at: http://www.oecd.org/dataoecd/42/47/42286309.pdf

Özden Ç, Parsons C, Schiff MW, Walmsley T. 2009. The Evolution of Global Bilateral Migration 1960-2000. In *The Second Conference on International Migration and Development, Sep 10-11*. The World Bank, Washington DC.

Parsons CR, Skeldon R, Walmsley TL, Winters LA. 2007. Quantifying International Migration: A Database of Bilateral Migrant Stocks. *World Bank Policy, Research working paper no. WPS 4165*.

Ratha D, Shaw W. 2007. South-South Migration and Remittances. *World Bank Working Paper No. 102*.

Santo Tomas PA, Summers LH, Clemens M. 2009. *Migrants Count: Five steps toward better migration data*. Washington DC: Center for Global Development.

Simmons AB. 1987. The United-Nations Recommendations and Data Efforts – International Migration Statistics. *International Migration Review* 21: 996-1016.

UNDESA Population Division. 2009a. International Migration Flows to and from Selected Countries: The 2008 Revision. *United Nations database, POP/DB/MIG/Flow/Rev.2008*.

UNDESA Population Division. 2009b. Trends in International Migrant Stock: The 2008 Revision. *United Nations database, POP/DB/MIG/Stock/Rev.2008*.

UNDESA Statistics Division. 1998. Recommendations on Statistics of International Migration: Revision 1. *Statistical Papers Series M* No. 58.

Vertovec S. 2007. Super-diversity and its implications. *Ethnic Racial Studies* 30: 1024-54.

Appendix: Expert Interviewees

Awad	Ibrahim	International Labour Organization
Batalova	Jeanne	Migration Policy Institute
Batgarjal	Uranbileg	World Bank
Bessudnov	Alexey	University of Oxford
Carvajal	Liliana	United Nations Development Programme
Castles	Stephen	University of Oxford / University of Sydney
Clemens	Michael	Centre for Global Development
Dumont	Jean Christophe	OECD
Diallo	Khassoum	UNHCR (email correspondence only)
Freinkman	Lev	World Bank
Goldin	Ian	University of Oxford
Hollifield	James	Southern Methodist University
Hovy	Bela	United Nations Population Division
Hugo	Graeme	University of Adelaide
Koehler	Jobst	International Organization for Migration
Kuznetsov	Yevgeny	World Bank
Martin	Philip	University of California, Davis
Newland	Kathleen	Migration Policy Institute
Özden	Çaglar	World Bank
Parsons	Chris	University of Nottingham
Passel	Geoffrey	Pew Hispanic Centre
Pereira	Isabel	United Nations Development Programme
Ratha	Dilip	World Bank
Skeldon	Ronald	University of Sussex
Swanson	Eric	World Bank
Walmsley	Terrie	University of Melbourne
Winters	Alan	University of Sussex / DFID
Yi-Ying Lin	Serena	Migration Policy Institute
Zetter	Roger	University of Oxford