Commercial airline travel and the international spread of emerging infectious diseases.

Thesis submitted in accordance with the requirements of The University of Liverpool for the degree of Doctor in Philosophy by Margaux Marie Isabelle Meslé.

29th June 2018

Table of Contents

Table of figures

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Chapter 6

Chapter 7

Appendices

Acknowledgments

The research was funded by the National Institute for Health Research Health Protection Research Unit (NIHR HPRU) in Emerging and Zoonotic Infections at University of Liverpool in partnership with Public Health England (PHE), in collaboration with Liverpool School of Tropical Medicine. Margaux Meslé is based at The University of Liverpool. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR, the Department of Health or Public Health England.

I would firstly like to acknowledge my supervisors, Dr Jonathan Read, Prof Rob Christley, Dr Ian Hall, Dr Steve Leach and Dr Roberto Vivancos for their support, mentoring, insights and patience, during these three years.

I would also like to thank my peers Dr Ben Francis, Thomas Inns, Dr Chris Jewell for their help in statistics, coding and modelling, respectively. You have all saved me many hours of work with your invaluable help. Maxine Martin and Vikki Grimbly from the administrative staff must also be thanked, for answering my many questions and for encouraging me with my work.

I must also thank my family (Meslé and soon-to-be Rowbotham) for cheering me on and supporting me along the way. Finally, a special thank you goes to Erik, for always making me smile and putting things into new perspectives, even after a long day. This one is for you!

List of acronyms

CAA: Civil Aviation Authorities

CDC: Centers for Disease Control and Prevention

EID: Emerging Infectious Diseases

GDP: Gross Domestic Product

HCS: HealthCare System

IATA: International Air Transport Association

IHR: International Health Regulations

IPS: International Passenger Survey

ONS: Office for National Statistics

PANYNJ: Port Authorities of New York and New Jersey

PHE: Public Health England

SARS: Severe Acute Respiratory Syndrome

TP: TravelPac

UK: United Kingdom, includes islands of Jersey and Guernsey

UN: United Nations

USDoT: United States Department of Transport

VBD: Vector-Borne Diseases

VFR: Visiting Friends and Relatives

WCS: Worst Case Scenario

WHO: World Health Organization

Note that OAG is no longer an acronym therefore has not been included as such.

“Commercial airline travel and the international spread of emerging infectious diseases” by Margaux Meslé

Abstract

A total of 1.186 billion international airline arrivals were recorded globally in 2015 alone, a 4.6% increase from 2014 (Glaesser *et al.*, 2017). As airplanes now fly very long distances at greater speeds, a passenger is likely to travel while incubating a pathogen and may only become ill once at their destination. In the 21st century alone, a number of pathogens have been transported in this way, causing epidemics (Cholera in Haiti, 2010) and pandemics (Influenza A H1N1, 2009). The aims of this thesis were to understand what previously purchased airline data represents in terms of passenger movement and whether this is a useful and/or accurate tool to use to predict the international spread of human infectious diseases.

A systematic literature review first analysed what airline data was most often used by mathematical modellers to determine the international spread of human infectious diseases and how well the data sources were reported. From there, the OAG airline data was extensively described and validated against independent and open access data sets. With a better understanding of the airline data, the author modelled which regions posed varied risks of chikungunya and dengue infection for UK passengers compared to the local populations, by combining endemic and imported number of cases to the airline data. Finally, the author conducted an analysis regarding which countries posed a higher risk for the initial spread of a pandemic by deriving their global connectivity from the airline data and using the level of healthcare provided from two indices. Both parameters were given equal importance by providing equal weightings before ranking each country by proximity to a fictitious ‘Worst Case Scenario’.

It was determined that commercial (closed access) airline data was most often used by the modelling community and that the reporting of sources used did not often allow for independent validation of a group’s work. As a result, a framework was developed for researchers to report specific aspects of the data set, such as date range included, any manipulation and date of collection. When describing the airline data, clear seasonal trends were apparent, and countries such as the United States and China contribute large numbers of passengers to the network. Additionally, the data are sold as highly accurate airline only data, but was identified as also containing land and sea transportation. When validated, the OAG data showed good agreement with the other data sets used such as from the United Kingdom’s Office for National Statistics and the United States’ Department of Transport. From the modelling chapters, some regions, such as the Caribbean, proved less dangerous for UK airline passengers in terms in chikungunya and dengue infection compared to the local populations whereas regions such as Lower South America were more dangerous for dengue specifically, for UK passengers. Using two independent indices and the same connectivity data, the author showed that certain countries exhibited the potential greatest risk of international pandemic spread, whereas countries with recent pandemic emergence, such as Brazil and Mexico, showed lower potential risk.

Future perspectives of this work include taking the global connectivity and healthcare chapter further by including within-country data. Additionally, the creation of an open-access data set combining detailed airline travel and passenger epidemiology that all research groups could use is an important continuation of this work.

Chapter 1 – Introduction

Preamble

This introductory chapter offers a brief overview of infectious disease events of note since the start of the 21st century, and the role played by airline travel in their international propagation. Additionally, a brief introduction to mathematical models is provided, with a short outline of what information models can provide policy makers and how to represent the airline network are also presented. Finally, the cost of pandemics is described before the aims of the thesis are listed.

Abstract

Within the 21st century alone, a number of infectious diseases outbreaks have spread rapidly across international boundaries, developing into pandemics, as a direct result of human airline travel. Outbreaks such as the 2009 H1N1 Influenza A pandemic or the Zika pandemic were the result of novel pathogens quickly spreading globally via human international movements. In contrast, the ongoing antibiotic resistance pandemic has resulted from decades of (usually inappropriate) antibiotic use allowing pathogens to evolve relatively unnoticed, but making their detection by the health authorities very slow. The high level of connectivity between countries means that no country is isolated and a pathogen can be transported into an epidemiologically suitable setting within a few hours, sparking a localised outbreak, as was the case in Haiti (2010) with cholera.

To gain a clearer understanding of the human airline travel patterns, mathematical models may provide crucial insights. Mathematical models are defined as a set of mathematical rules designed to represent a biological system, in a more or less complex manner, based on a set of parameters. Models not only help understand the epidemiology of a pathogen, but also provide an understanding of the consequences of implementing a given policy. When based on the airline network, mathematical models help understand the possible development of an outbreak into a pandemic and geographical spread. The network that is made up of airline movements is known to be a 'small-world' network, in which any two nodes (here airports) can be reached in a given small number of steps (here routes), compared to that expected in an equivalent random network.

The origin of a pandemic is known to be unpredictable (both in time and geography) and the resulting outbreak may be very costly (in fatalities and economic burden) for the affected countries and the global community. However, by using mathematical models, these costs can be reduced, through a better understanding of a pathogen's potential spread.

Therefore, through a thorough analysis of the global airline network, this thesis aims to determine whether the use of airline data were appropriate to understand the international spread of human infectious diseases, through four main objectives.

Introduction

In March 2018 the first direct flight between London (United Kingdom) and Perth (Australia) successfully landed after 17 hours flying time (BBC, 2018). In the last century, this trip would have taken at least 24 hours, and one month the century before that (Cliff and Haggett, 2004). Important advances in transport technology over recent centuries have improved how distant cities and populations are connected to each other, with the largest cities becoming increasingly connected, and previous refuelling points (for airplanes and ships) becoming less required as airplanes now need fewer stops for equal distance travelled (Cliff and Haggett, 2004).

An increasing number of passengers travel annually, with 1.186 billion international arrivals recorded globally in 2015, a 4.6% increase from the previous year (Glaesser *et al.*, 2017). This increasing number of airline passengers is posing an increased burden on international public health organisations. Indeed one of the biggest threats faced by modern populations globally is the speed, distance and number of airline passengers travelling today, the likes of which have never previously been seen (Lopez *et al.*, 2016). Healthcare practitioners need to correctly diagnose and treat for pathogens they may be unfamiliar with, all while trying to avoid an outbreak (Fricker and Steffen, 2008). Passengers are faced with varying levels of risk when travelling, with a small proportion of them making more frequent trips and mixing with similar people (so called assortative, or like-with-like, mixing). These frequent flyers are more likely to board an international flight while in the incubation or asymptomatic phases of an illness (Hollingsworth *et al.*, 2007), increasing the potential of introducing a pathogen into a susceptible population, potentially resulting in an outbreak (Tian *et al.*, 2017).

International travel has been determined to be an important driver of novel or Emerging Infectious Diseases (EID) (Arcilla *et al.*, 2017; Fricker and Steffen, 2008; Semenza *et al.*, 2016a). According to Jones *et al.* (2008), EIDs can be defined in a number of ways, such as 1) novel pathogens or mutated strains of known pathogens (Multi-Drug Resistant Tuberculosis, for example); 2) pathogens entering naïve populations or ones with reduced immunity, such as Ebola; or 3) those showing a recent increase in incidence in a given population (measles in Europe (World Health Organization, 2017c)). Pathogens such as the Human Immunodeficiency Virus (HIV) affect the host's immune system leading to an increased susceptibility to other pathogens, as was seen at the start of the pandemic in the 1980s and 1990s (Jones *et al.*, 2008). If the necessary protective precautions are not adhered to correctly, travellers visiting tropical and subtropical countries pose a health risk to themselves as well as to local populations, and populations from their home country, as they can transmit a pathogen to local vectors in the visited country (Semenza *et al.*, 2016a) or in their home country (Angelini R *et al.*, 2007; Mier *et al.*, 2017). An example of passengers introducing vector-borne pathogens to local vectors upon their return from international travel was the 2007 chikungunya outbreak in the Ravenna region of Italy (Angelini R *et al.*, 2007).

Monitoring the levels of disease importation resulting from travel may help determine epidemiological changes in visited countries (Fricker and Steffen, 2008; Lopez *et al.*, 2016), which along with monitoring underlying drivers of emerging pathogens may help identify early cases, which may accelerate and improve outbreak detection and control (Semenza *et al.*, 2016a).

Human pathogens follow human travel patterns and are therefore more likely to start spreading within the same region as the majority of passengers (four out of five) travel within their own region (Glaesser *et al.*, 2017). In an effort to control outbreaks early, the World Health Organization (WHO) coordinates the international response to reported outbreaks from its 194 member states adhering to the International Health Regulations (IHR). This set of guidelines updated in 2005 after the Severe Acute Respiratory Syndrome (SARS) outbreak, legally requires countries to report cases or outbreaks that may require international coordination especially in the event of an international spread. Additionally, sentinel organisations such as GeoSentinel and Trop Net Europe collect disease importation information from general practitioners to help understand the international importation risks (Fricker and Steffen, 2008).

Human infectious diseases events of importance in the 21st century

Since the start of the 21st century, a number of infectious disease events have taken place globally, with at least four causing international concern: SARS (2003), H1N1 Influenza A 2009 pandemic strain, Ebola virus (2014) and Zika virus (2016). This situation is at odds with what some in the medical community believed about infectious diseases in the 1970s, when it was thought the war on micro-organisms was all but over. However, important factors such as pathogen evolution, ecology and human travel were not considered at the time as posing significant threats (Arnal *et al.*, 2011). The next section will describe outbreaks of importance from the 21st century alone and how unique influential factors shaped each outbreak's severity, ranging from pandemics to significant localised outbreaks with pandemic potential.

Pandemics

Infectious disease pandemics refer to outbreaks that have affected populations over a large geographic area, countries or continent (Porta, 2008). These outbreaks may cause fear in populations leading to behaviour changes and population movements that in turn cause important economic loss and/or increased mortality rates (International Working Group on Financing Preparedness, 2017).

An unusual rise in infectious disease cases of any kind needs to be detected and controlled very early on to allow an appropriate response and resource allocation. Early detection and control depends significantly on good communication within governmental branches and with the private sector (International Working Group on Financing Preparedness, 2017). This absence of communication within governmental authorities led to an important time lag allowing the SARS virus to spread widely within Southern China (Bowen Jr and Laroe, 2006; World Health Organization, 2003). The pandemic started in a live market in Guangdong province, Southern China, probably in November 2002. A then novel coronavirus that emerged to cause the 2003 pandemic is now known to have a reservoir in three animal species (civet cats, badgers and dogs) (Brower, 2003). Even though a number of super-spreading events (one infection event creating a much larger than average number of next generation events) occurred during the outbreak and no cure was available, the disease was controlled more easily than others as patients were only infectious when symptomatic (Heymann *et al.*, 2013). For example, in February 2003, a health care professional travelled to Hong Kong and was taken ill when staying at an international hotel, leading to a number of infections (Bowen Jr and Laroe, 2006; World Health Organization, 2003) as susceptible clients came into contact with virus shed in the hotel corridor and lift by the health care worker (Bowen Jr and Laroe, 2006). Within ten days of the virus being introduced in Hong Kong, cases were reported across the world, in Canada, Vietnam, and the United Kingdom (UK), among others, as the direct result of airline passengers travelling internationally (Bowen Jr and Laroe, 2006). After this international spread, the WHO advised airport staff to question passengers about symptoms when leaving the airport from March 2003, and the following month advised against all non-emergency flights to affected areas. By May, air travel to and from China fell drastically. The outbreak was declared over in July 2003, after a total of 8,096 cases, including 774 recorded deaths (Bowen Jr and Laroe, 2006). Although air travel restrictions played a role in stopping the international spread of the virus through airline passengers, the economic impact on the airline and tourism industries was significant. According to the International Working Group on Financing Preparedness (2017) China alone is thought to have lost 0.5% of its Gross Domestic Product (GDP) as a consequence of the travel advisory. Great international collaboration and cooperation allowed the outbreak to be brought under control and person-to-person transmission stopped within eight months (Heymann *et al.*, 2013).

In response to the significant economic losses from the airline travel advisory and initial slow case reporting to and by the Chinese health authorities, WHO revised its IHR policies to encourage faster detection and reporting by countries. The new guidelines came into effect in 2007 (Heymann *et al.*, 2013), four years before the next pandemic occurred, the 2009 H1N1 Influenza A pandemic. The novel strain had been shown to infect healthy Mexican populations, causing severe disease outside of influenza season (European Centre for Disease Prevention and Control, 2010). As a result, case reporting for the novel Influenza A virus strain in two Californian children was reported to the Centre for Disease Prevention and Control (CDC) early after detection. Four days later, on the 25th of April, WHO declared the outbreak a “public health emergency of international concern” (European Centre for Disease Prevention and Control, 2010). However, even with early detecting and reporting of the novel strain, WHO thought that the virus had already spread too far internationally and therefore advised against airline restrictions. Indeed given the virus natural history (being infectious before being symptomatic) makes its control more difficult, and the virus was reported on four continents within three weeks of Mexican authorities reporting the outbreak, most likely as a result of international air travel and trade. It was therefore too late for an airline ban to have any significant impact on the development of the pandemic (Hosseini *et al.*, 2010). Despite its close genetic proximity to the virulent 1918 strain (European Centre for Disease

Prevention and Control, 2010), the 2009 pandemic strain was much less virulent than feared due to previous immunity in older generations (leading to fewer mortalities) but a high mortality among children and young adults was recorded (Fineberg, 2014). The outbreak was officially declared over on the 10th of August 2010 after 68 weeks, 925,861 cases (European Centre for Disease Prevention and Control, 2010) and 18,449 deaths reported (World Health Organization, 2010), although some researchers argue this may be a severe underestimation of the true number of cases and deaths (Dawood *et al.*, 2012).

Although the implementation of the revised IHR guidelines showed a positive impact on the H1N1 pandemic, such as improved communication between WHO and member states, it became apparent that many countries were unprepared to handle a future pandemic (Fineberg, 2014).

Contrary to the previous two pandemics for which the spread was well reported and described in the media, Zika virus spread slowly from Africa through South Asia and into the Americas without causing international concern until it was detected in Brazil in late 2015 (Basundra *et al.*, 2016). During its eastward spread from Africa, the virus caused small, localised outbreaks with symptoms similar to those of dengue virus infections (Jamil *et al.*, 2016). Although the virus mutated over time, it still caused low mortality, but an increasing trend in new-born microcephaly and adults presenting with Guillain-Barré syndromes became more prominent (Chang *et al.*, 2016; Mlacker *et al.*, 2016). The introduction of the virus to the American continent a few months prior to the 2016 Rio de Janeiro Olympic games may have provided a pathway for the virus to be transmitted further on the continent via airline travel (Chang *et al.*, 2016). Overall, the outbreak is estimated to have cost between US$2.3-6 billion per year, or 0.05-0.12% of the global GDP. Such costs, partly due to the surprise and timing of the outbreak, could have been mitigated by more sensitive and stronger healthcare systems. These could have been in a position to disseminate accurate information about co-factors for microcephaly, transmission patterns and disease understanding (United Nations Development Programme, 2017). It is likely that once again airline travel bans for passengers travelling to and from the American continent would not have been beneficial (and would have resulted in bigger economic losses) as these would have been implemented too late to have proved any benefit for the spread of the virus.

Antibiotic resistance importation

The pandemics described so far have had significant costs associated to them, both economical and in terms of mortality or morbidity. However, viruses are not the only pathogens being transported by human travel (Amesh *et al.*, 2018). In fact, an emerging and important public health problem therefore is the international movement of antibiotic resistant bacteria.

The emergence of a drug resistant bacterial strain in a given location should not be considered an isolated event, but rather a global threat, given the global airline passenger volume and level of connectivity (Choudhury *et al.*, 2012) and posing a significant threat to the entire modern medicine (Amesh *et al.*, 2018). Several examples of imported drug resistant bacteria have been reported in Spain and Sweden after travellers and diaspora returned to their country of residence (Choudhury *et al.*, 2012). They may acquire the pathogen and be asymptomatic in these communities for an extended period of time after their return. As Arcilla *et al.* (2017) show, 11.3% of passengers returning from international travel were still colonised with Extended Spectrum β-Lactamase-producing *Enterobacteracae* (ESBL-E) one year after their return. This result was accompanied by a 12% risk of onward transmission within the household (Arcilla *et al.*, 2017).

Several antibiotic resistant strains have been identified as imported from several global regions, with varying levels of associated risks. However, India has been noted to cause the largest risk to international travellers regarding antibiotic resistant bacterial strains (Choudhury *et al.*, 2012). Two resistant *Escherichia coli* (*E. coli*) strains (*enterotoxigenic* and *enteroaggregative*) have been identified in returning travellers from India, as was also the case for the New Delhi Metallo-betalactamase 1 resistance gene discovered in Swedish diaspora in 2007 (Choudhury *et al.*, 2012). Asia was reported as the most likely region to acquire ESBL-E resistance gene (Arcilla *et al.*, 2017; Choudhury *et al.*, 2012), followed by Central and Eastern Asia (Choudhury *et al.*, 2012). Pre-travel advice on reducing the use of antibiotics could play an important role in reducing the importation of ESBL-E genes by travellers (Choudhury *et al.*, 2012), as well as other resistance genes or pathogens. The antibiotic resistance pandemic is estimated to cost the global economy up to 3.8% of the global GDP by 2050 and to impact low-income countries most. This cost is higher than the predicted cost of a pandemic, estimated at up to 1% of the global GDP (International Working Group on Financing Preparedness, 2017). As well as this significant financial cost, the morbidity and mortality rates are going to rise (World Bank, 2017a). These costs will increase rapidly if the spread of resistant strains is undetected as a result of weak surveillance, delayed appropriate treatments to be provided to human as well as animal cases. This in turn puts additional strain on control measures as they will be confronted with a higher number of cases (World Bank, 2017a).

Recent localised outbreaks

Within the past decade, airline passengers travelling between distant countries have imported pathogens, resulting in country-level outbreaks of importance within the international destination. Examples include the importation of *Vibrio cholerae* to Haiti (2010) and the Middle East Respiratory Syndrome Corona virus (MERS-CoV) to South Korea (2015). Two additional outbreaks of note that developed although without airline importation but saw a number of air-travel associated exports were the West African Ebola (2014-2015) (classified as a "public health emergency of international concern" by WHO) and the Madagascar plague outbreak (2017) (World Health Organization, 2016b).

On the 21st of October 2010, the Haitian Ministry of Public Health declared a cholera outbreak of particular virulence. This was the first reported cholera outbreak on the island for over a century. The resulting fear in the local population led to populations moving from Meille, (the epicentre) which allowed the bacteria to be carried to distant locations across the country (Piarroux *et al.*, 2011). The introduction of a hyper-virulent strain with high infectiousness in a naïve population made this outbreak particularly virulent (Piarroux *et al.*, 2011). Additionally, the outbreak occurred a few months after the devastating earthquake had left the capital, Port-au-Prince, in a fragile state with 1.3 million people homeless and in makeshift camps throughout the capital as well as thousands of deaths and injuries. Within these precarious living conditions, with little sanitation and access to water, an infectious disease outbreak was reported as highly likely to develop (Walton and Ivers, 2011). The introduction of the bacterium (*Vibrio cholerae O1*, serotype Ogawa, biotype El Tor) was linked to an outbreak in Kathmandu (Nepal) that occurred just one month prior to the one in Haiti. The link between the two outbreaks was quickly established and the pathogen's entry into Haiti was linked to a Nepal-based battalion suffering from cholera, who arrived a few days prior to the first cases being reported locally (Chin CS, 2011). As of the 30th of November 2011, a total of 515,699 cases were reported of which 54% were hospitalised and 1% died, according to Pan American Health Organisation (2011). This outbreak was still ongoing at time of writing, with cases reported on a regular basis (United Nations News Centre, 2017). This unusual outbreak resulting from a single importation event but contained to one country is a reminder that a disease outbreak in one country can very quickly affect populations much further away, and highlights the importance of the IHR and rapid reporting of cases.

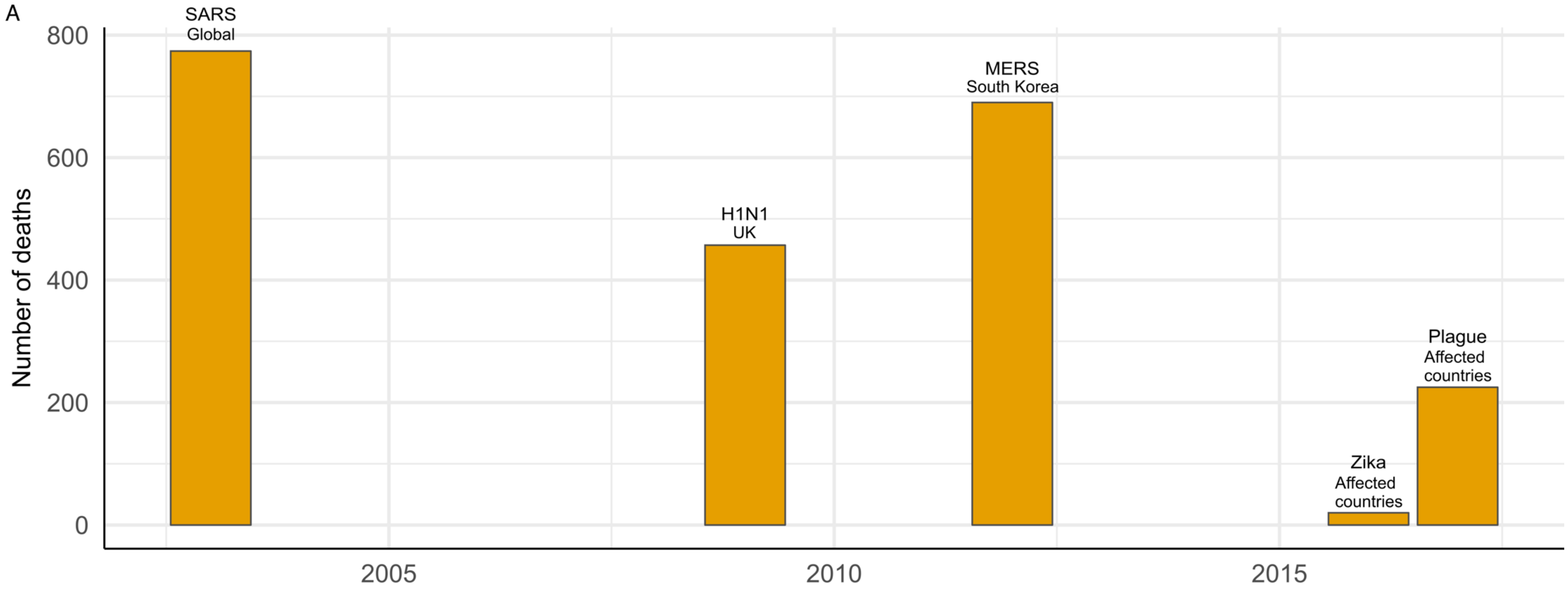
Although first identified in the Middle East in 2012 as a human respiratory virus, the MERS Corona virus is now known to have circulated in dromedary camels for decades prior to causing human disease (Gardner *et al.*, 2016). Human-to-human transmission is uncommon except in close contact and hospital settings, with sporadic cases linked to close proximity with dromedary camels. The majority of cases (80%) have been concentrated in the Middle East, with airline passengers exporting the virus to 27 countries, but rarely causing an outbreak (World Health Organization, 2017d). As of March 2018, there have been a total of 2,144 laboratory confirmed cases of MERS, of which 750 were fatal (World Health Organization, 2018c). Few exported cases have led to an outbreak, of which the most notable was in South Korea during the summer of 2015, when a patient became ill with the virus upon his return from a business trip in Saudi Arabia (Su *et al.*, 2015). After visiting three hospitals, the patient (who was later determined to be a super-spreader) caused an outbreak of 186 cases, due to slow identification of the virus by healthcare professionals (Gardner *et al.*, 2016; Ki, 2015). Because the Middle East is very well connected to the rest of the world, millions of airline passengers travel through airports such as Dubai and Doha annually. Along with mass gathering events, such as the Hajj, taking place every year there is a clear risk of further international dissemination of the virus from this region by international airline travel (Gardner *et al.*, 2016; Al-Tawfiq *et al.*, 2014). Even though the risk of onward international transmission is high for this ongoing outbreak, no flight or travel restrictions have been put in place by WHO at time of writing (World Health Organization, 2017d).

These two examples (cholera and MERS-CoV) illustrate the potentially dramatic effects of a small number of airline passengers may have on local populations, the next examples highlight the potential risk of international dissemination from an outbreak.

The largest Ebola virus outbreak ever recorded occurred between 2014 and 2016 in West Africa, specifically Liberia, Guinea and Sierra Leone (Dhama *et al.*, 2015) with few exported cases, causing a total of 28,616 confirmed cases and 11,310 deaths (World Health Organization, 2016b). Although the outbreak principally affected the three previously named countries, onward transmission from exported cases was also recorded, although to a smaller extent and mostly in Nigeria (20 cases), Mali (8 cases) and United States (4 cases) (Elmahdawy *et al.*, 2017). It was determined that the burial practices of the local populations facilitated transmission between family members and close contacts as the virus is transmitted by bodily fluids (Dhama *et al.*, 2015). An important challenge faced during the outbreak was keeping track of cases and their numbers, as many could not attend a hospital because of distance, overcrowding and/or stigmatisation (Elmahdawy *et al.*, 2017). Prior to the outbreak, the healthcare facilities in these countries were far from end-users, closed (due to lack of staff or equipment) or overcrowded. Additionally, the population had a strong belief that hospitals did not heal patients (but rather killed them) causing many families to hide cases, allowing the virus to spread further (Omoleke *et al.*, 2016). The three West African countries had seen a recent and important post-conflict economic growth, leading to improvements in transportation (including international airline travel). Indeed, some 39 weekly inter-continental flights from the three capitals with major international airlines were available prior to the outbreak, but the healthcare system still remained vulnerable. These flights were suspended during the outbreak due to the perceived high risk of international dissemination of the virus (Omoleke *et al.*, 2016) against the advice of the WHO (Nutall, 2014). Quick international transport links made the potential for the international dissemination of cases at the start of the outbreak a high risk (Dhama *et al.*, 2015). The WHO lifted the state of "public health emergency of international concern" on the 29th of March 2016 (World Health Organization, 2016b).

Another example of the importance of global connectivity was seen in Madagascar. In 2017, the island saw an unusually large outbreak of plague recording 2,417 cases including 209 deaths recorded by the 26th of November 2017. The peculiarity of this outbreak was that 77% of cases were clinically pneumonic rather than bubonic (World Health Organization Africa, 2017) and affected urban rather than rural areas (Burki, 2017). The risk of international dissemination through airlines has been assessed as low by WHO, but a number of closely connected countries (Seychelles) put a temporary flight ban in place (World Health Organization, 2017f) or set up preparedness actions (South Africa, Tanzania, among others) (World Health Organization, 2017f). Bogoch *et al.* (2018) analysed which countries were most at risk of an importation event, potentially resulting into an outbreak, using each nation's airline connectivity with Madagascar and their health care system capacity to cope in the event of an outbreak. Although no international exportation events have taken place from this outbreak, this is one of the latest outbreaks with the potential to cause international concern at time of writing (the Nipah virus outbreak in India and the Ebola outbreak in the Democratic Republic of the Congo should also be considered). Madagascar's healthcare system had deteriorated since the 2009 coup and is now one of the most underfunded systems globally, as well as being ineligible for international aid. This outbreak highlighted the need for quick and accurate detection of pathogens by the local healthcare systems (Bonds *et al.*, 2018). Given the situation with a possible risk of international dissemination, WHO helped the nation by providing personal and financial support (Bogoch *et al.*, 2018). Although treatable with antibiotics, the disease is still feared (World Health Organization, 2017e).

The examples highlighted above show the potential impact of rapid airline travel on global health, whether these importation events occurred rapidly and were easily noticed, such as the H1N1 outbreak, or slower and more difficult to identify, such as antibiotic resistant bacteria. Pandemics are also known to be very costly to the global economy, as well as that of the affected countries. For example, the 2003 SARS outbreak cost an estimated US$52.2 billion to the global economy (International Working Group on Financing Preparedness, 2017). This significant economic impact is the result of poor pandemic preparedness within countries as well as changes in population behaviour. In fact, it was noted that the important cost of MERS-CoV on the South Korean economy (US$ 10 billion) was in part a result of flight cancellations by travellers from fear of becoming infected when visiting the country (International Working Group on Financing Preparedness, 2017). As well as having a significant impact on local and global economies, pandemics can also cause a large number of deaths in the affected countries (International Working Group on Financing Preparedness, 2017), even if the case fatality rates vary according to the pathogen, as shown in **Figure 1.1**. Both of these costs can be significantly reduced by improving country level preparedness, which must be taken both nationally with appropriate reference to healthcare systems (including trained staff, appropriate facilities and trust from the population) and globally with international surveillance systems that report outbreaks accurately and in a timely manner (Omoleke *et al.*, 2016). Therefore, understanding the dynamics and spread of future outbreaks and predicting origins, destinations and speed of dissemination is a research field of its own that, through mathematical modelling can help public health organisations make better and more informed decisions concerning the development of an outbreak into a pandemic and containment measures.



(Figure 1.1 continues on next page)

(Figure 1.1 continued)

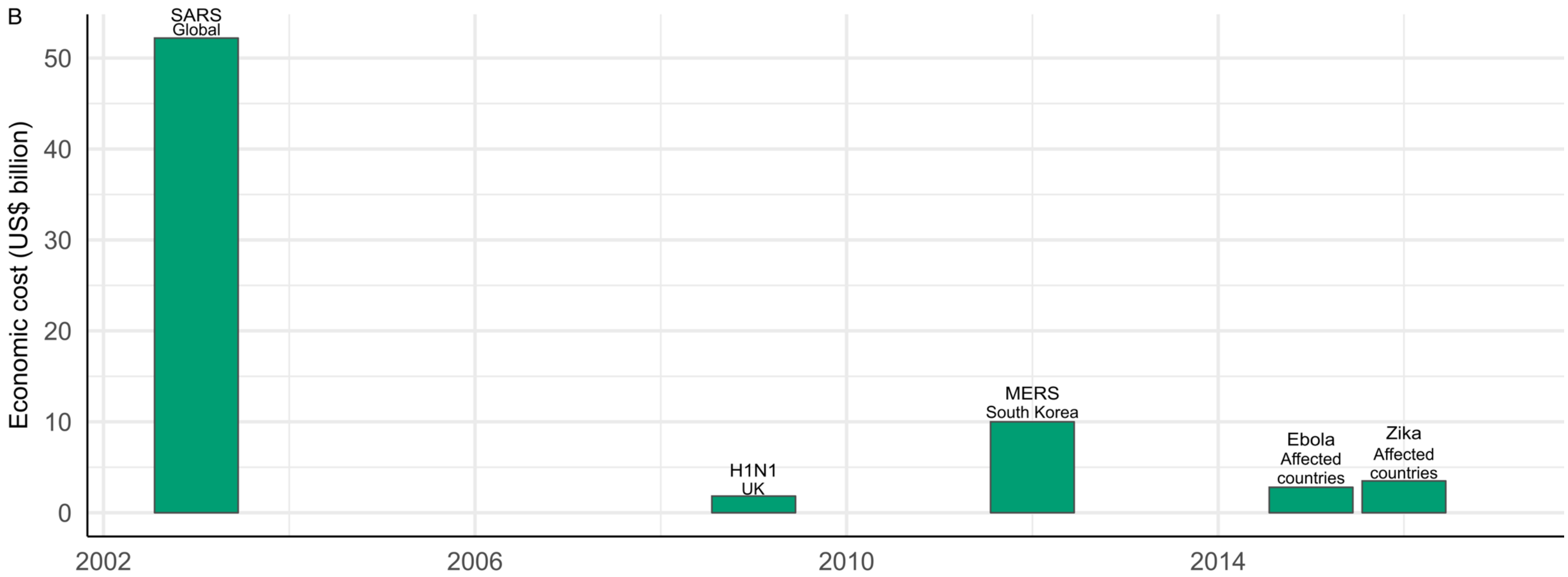


Figure 1.1: A) Total number of deaths and B) total economic cost (in US$ billion), resulting from the four pandemics of the 21st century by region or country affected.
Data from: International Working Group on Financial Preparedness (2017); Hine, H. (2010) The 2009 Infuenza Pandemic - an independent review of the UK response to the 2009 influenza pandemic; Pan American Health Organization (2018).
Note: the H1N1 Influenza A pandemic affected more than just the UK, however, this country was used as an example.

Introduction to mathematical models

When studying the transmission cycle of malaria in the early 20th century, Ronald Ross derived a series of equations relating to the parasite's transmission cycle and potential effects of vector control on malaria incidence in the local population. Still relevant today, his work is one of the earliest examples of models using mathematical equations to understand the spread of human infectious diseases (Anderson and May, 1992; Lessler and Cummings, 2016). Contemporary with Ronald Ross was William Hamer who, while studying measles epidemics, understood that epidemics depend on the contact between infectious and susceptible parts of the affected population (Anderson and May, 1992). As well as being among the first to use mathematical equations to represent infectious disease spread, Ross and Hamer also worked on understanding the parameters used within those equations (Anderson and May, 1992). Later on, in the 1930 and 40s, Reed and Frost taught epidemiological theory at John's Hopkins University using mathematical models and mechanical epidemiology simulators. These are some of the earliest examples of the use of mathematical equations to understand the spread and control of infectious diseases (Lessler and Cummings, 2016). Later, in the 1980s, Rvachev and Longini developed the first model to explore the international spread of influenza using airline data from 52 airports located on all continents (Rvachev and Longini Jr, 1985).

Mathematical models are useful tools to understand the spread of infectious diseases within and between populations and the potential impact of policies on their spread and control (Lessler and Cummings, 2016), as well as understanding a pathogen's mechanical (within a host) and epidemiological (between hosts) spread. In today's increasingly connected world, and given the unpredictability of outbreaks, understanding and knowing how to control an outbreak from an unknown imported case is crucial to avoid the further spread of an infection (Hollingsworth *et al.*, 2007).

The next section will introduce the concept of mathematical models, why these are important and what previous models of international infectious diseases spread have shown.

What are mathematical models

Mathematical models are conceptual tools used to explain the behaviour of objects (in this context, objects being humans and pathogens) (Arnal *et al.*, 2011) in a set of precise mathematical rules, thus providing a clear and concise language (Huppert and Katriel, 2013). In public health, their use ranges from understanding the epidemiology of an infection within a host, to understanding of the impact of vaccination on infectious disease spread within populations. Models are more or less complex (Lessler and Cummings, 2016), but are all wrong on some level, as simplifications and assumptions must be taken at some stage of developing the model (Keeling and Rohani, 2008). These simplifications may affect any part of the model, such as using a unique level of infectiousness between members of a population (ignoring super-spreaders), or assuming that the population as a whole is at equal risk of infection (ignoring previous immunity) (Ferguson *et al.*, 2003). Models help predict the course of an outbreak and guide policy makers in making difficult decisions, but also help understand how a pathogen spreads within a population and any influencing dynamics (Basu and Andrews, 2013; Keeling and Rohani, 2008; Lessler and Cummings, 2016). Each factor can be examined independently, creating an ideal world for disease study. However, models can never be fully accurate as some unknown behaviours will always be present in populations and disease transmission, this uncertainty in turn provides the confidence intervals for the model and the levels of risks among different groups (Keeling and Rohani, 2008). The development and availability of tools to predict and evaluate the impact of varying policies affecting the spread of infectious diseases is important. However, these tools must also be flexible enough to represent different populations and pathogens (Arnal *et al.*, 2011).

An ideal model should be based on essential features, and must balance a number of essential elements, namely 1) accuracy (data availability, computer power and understanding of the disease will determine a model's complexity); 2) transparency (using different elements in turn to understand their role, but considering that more parameters make for more complicated models) and 3) flexibility (how adaptable the model is to another pathogen) (Basu and Andrews, 2013; Keeling and Rohani, 2008). An oversimplified model will lead to the wrong conclusions whereas an overcomplicated model will obscure clear understanding of the results (Basu and Andrews, 2013; Ferguson *et al.*, 2003; Grassly and Fraser, 2008). Researchers must also be able to parameterise the model with the data available to them, which may create a difficult research situation when considering an emerging pathogen with little data available (Ferguson *et al.*, 2003; Keeling and Rohani, 2008). Models should be validated to ensure they correctly represent the disease dynamics where possible, and this should ideally be done using independent data or by using good statistical methods. Models used to inform policy must capture the underlying mechanisms of the policy being considered (Ferguson *et al.*, 2003).

Different model types have been developed over time to understand the spread of infectious diseases, with Kermack and McKendrick (1927) being among the earliest to describe an epidemic through a contact network, by developing the Susceptible-Infected-Recovered (SIR) model for a given pathogen in a population of size N (**Figure 1.2**). SIR models represent the pathway of how individuals start as susceptible (no previous infection), become infected (have acquired the infection), and recover (either acquire immunity or become susceptible again) (Arnal *et al.*, 2011; Keeling and Eames, 2005). A number of variations of this simple model exist, taking into consideration, for example, the population as a whole (including number of births and deaths, previous immunity), as well as whether the patient is infectious before becoming symptomatic (latency period) (Keeling and Rohani, 2008). These models and variations there-of have provided important knowledge on the epidemiology of a number of human pathogens, even with a number of significant assumptions having to be made. Some of the most important assumption for these models are that all members of the population mix homogeneously, and therefore are at equal risk of infection from each other and that once infected, the level of individual infectiousness remains constant over time and is the same for all infected individuals (Ferguson *et al.*, 2003). These, of course, cannot be true in a real-world population, as previous infections and/or vaccinations will alter the number of truly susceptible individuals in a population and different individuals will interact with other members of the population in different ways. Therefore, including these assumptions in a model will significantly impact the results, especially if those assumptions are invalid in the context of the specific use of the model (Ferguson *et al.*, 2003; Grassly and Fraser, 2008).

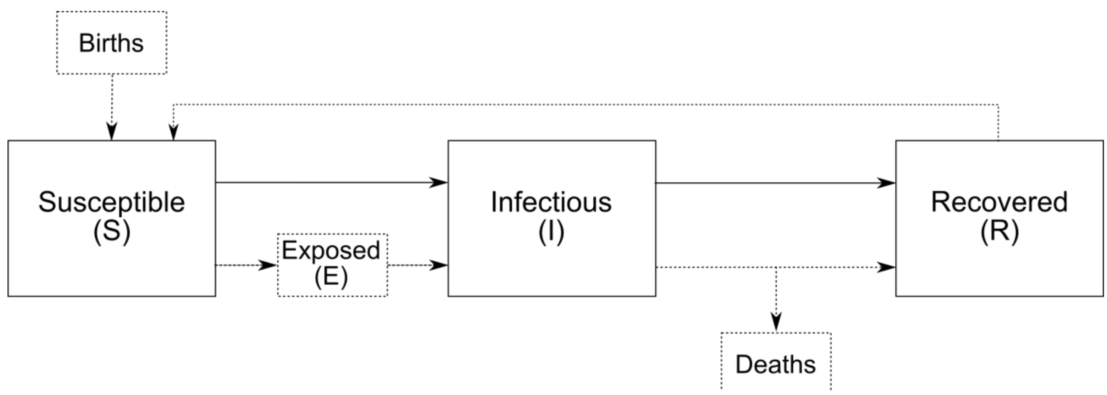


Figure 1.2: Example of an SIR model, with possible variations in the dashed boxes and lines.

When considering the spread of infectious disease between multiple international populations, assuming homogenous population mixing may not be correct. Therefore, it may be more appropriate to use an alternative approach such as a network representation of connectivity between individuals (Keeling and Eames, 2005), as the contact patterns within a population network are heterogeneous and each individual will only come into contact with a small fraction of the total population (Keeling and Rohani, 2008). Although they may be more difficult to interpret, network models are very important when contacts occur within a heterogeneous population (Arnal *et al.*, 2011).

A network represents a set of nodes (airports or countries for example) joined by links, also called contacts (direct flight routes for example), as a function of time. In the case of the airline network, two airports are linked if there is a direct flight between them (Arnal *et al.*, 2011). A network can also be directed (also called weighted or asymmetric, if for example more passengers travel towards one airport and then from it) or otherwise be undirected (also called unweighted or symmetric, if passengers travel in both directions between both airports in roughly equal numbers). Additionally, a network's links may be weighted, by using for example, a proportion of the number of airline passengers travelling between airports (Barrat *et al.*, 2014). Other ways of describing a network include the 'geodesic path', defined as the shortest path between node pairs across the network, with the 'diameter' being the longest shortest path between any two nodes across the network (Newman, 2003). Some additional characteristics will be described in **Chapter 3**.

The global airline network is an example of a small-world network (Barrat *et al.*, 2014; Guimera *et al.*, 2005; Wandelt and Sun, 2015), where each airport is connected to many neighbouring airports, with a few being additionally connected to distantly located airports (**Figure 1.3**) (Newman, 2003). This provides high levels of clustering in terms of number of neighbours (average probability that two nodes are in neighbouring networks are also geographic neighbours (Newman, 2003)) and little heterogeneity, allowing an infection to spread quickly within a cluster, with the long-range connections providing an outlet for the infection to reach other parts of the network quickly (Watts and Strogatz, 1998). There are two main characteristics of small-world networks, namely that there are neighbourhoods within the nodes, and the network diameter increases logarithmically with the number of nodes. The second characteristic allows all nodes within the network to be connected to each other in a small number of steps, meaning that the network has a finite dimension (Amaral *et al.*, 2000).

Understanding a network's topology is important, however, so is understanding its vulnerabilities. In general, the removal of at least one node from a network is enough to cause its breakdown. Therefore, the elimination of a central airport will result in more important consequences to the network than by eliminating a single random airport. This is a direct result of the overload generated on the rest of the network by altering traffic flows (Tran and Namatame, 2016).

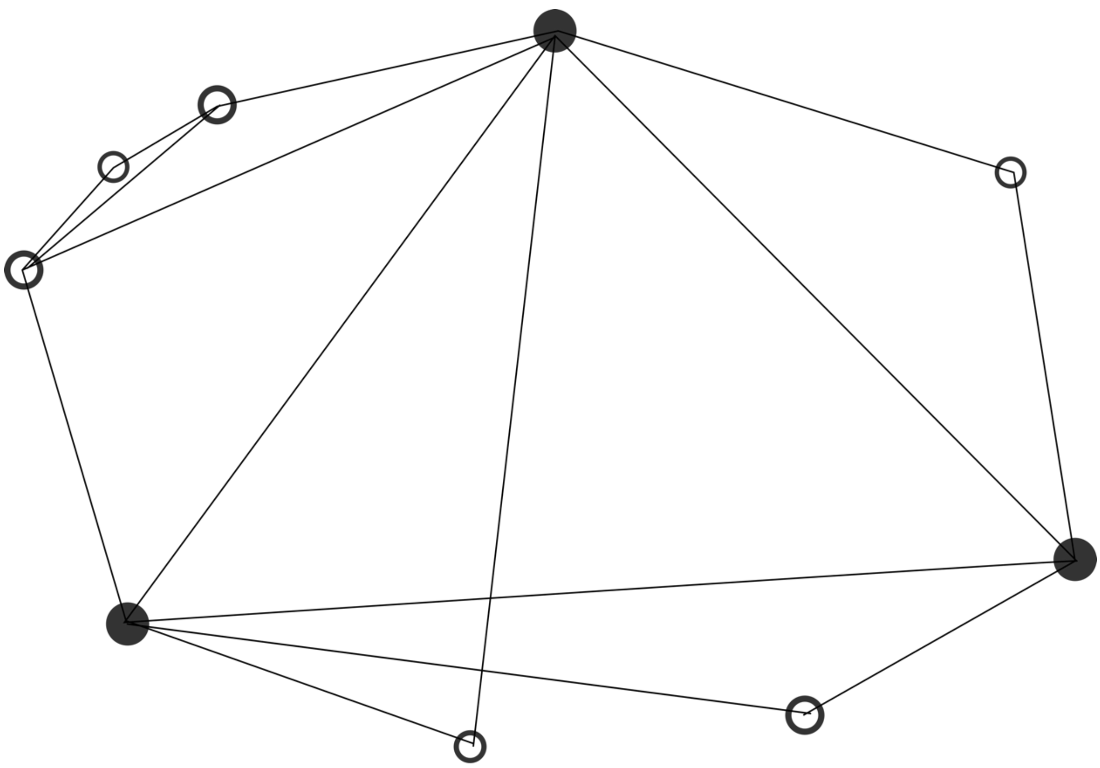


Figure 1.3: Example of a small world network representing the global airline network. Circles represent airports (nodes), with hubs represented as full cirles and the links represent the flow of passengers in both directions.

What are models used for and why are they important?

If developed correctly, with the appropriate data and methodology, mathematical models are powerful tools that help understand a pathogen's mechanics, potential spread and effective control measures (Alshammari and Mikler, 2016; Arnal *et al.*, 2011; Lessler and Cummings, 2016). Models help inform public health policy in various ways, ranging from broad information on feasibility to detailed recommendations, allowing policy makers to choose between interventions and compare investments (Lessler and Cummings, 2016). The aim of such models was to determine which individuals will be most infectious in order to reduce transmission (vaccination, isolation, treatment) and control the outbreak more efficiently (Grassly and Fraser, 2008).

Models can help understand the basic epidemiology of a pathogen and its control (Keeling and Rohani, 2008; Lessler and Cummings, 2016). This may be of particular relevance for a novel or re-emerging pathogen, especially in a naïve population, to limit the size of the outbreak early on, thereby reducing economic costs, and health impact. Building on a better understanding of basic epidemiological principles, models can also help understand surveillance (Lessler and Cummings, 2016). For example, by using surveillance data (known number of cases) and the known incubation period, Ron Brookmeyer was able to understand the true number of cases infected by Human Immunodeficiency Virus (HIV) (Brookmeyer, R. 2016). In other words, he calculated the number of unreported cases. From this, he was then able to predict how many HIV infections would then develop to Acquired Immune Deficiency Syndrome (AIDS) (Lessler and Cummings, 2016).

Models allow researchers to understand the potential role of interventions without the use of experimental epidemiological studies as these may not be feasible in the population (for cost, ethical or other reasons). Although challenging, forecasting and preparing for catastrophic events (even if unlikely to occur) are crucial for population health and security. For example, when the next influenza pandemic emerges, understanding the timing of antiviral drug distribution and the potential impact of social-distancing measures (such as school closures) and whether this would suffice to mitigate the outbreak may prove crucial in reducing the number of cases (Lessler and Cummings, 2016). Indeed, once a model has been formulated and calibrated; it can be used to understand a potential future outbreak. For example, a model can help determine the number of potential cases, how these numbers may vary with vaccination rates, estimate the epidemic curve, among other useful pieces of information (Huppert and Katriel, 2013). Such a model was implemented by Klepac *et al.*

(2018), modelling an influenza pandemic in the United Kingdom using mobile phone data. The model considered the introduction of targeted vaccination and hand hygiene as control measures and their impact on the spread of the outbreak. The results showed the significant impact these control measures had on the spread of the outbreak, both in terms of lives saved and significant reduction in speed of spread. Such models are likely to influence policy makers as to which control measures to prioritize when the next pandemic is detected in the country (Klepac *et al.*, 2018). Finally models can help understand the potential consequences of policy changes on one or more pathogen (Lessler and Cummings, 2016). It is important to not only report the worst and most likely scenarios to policy makers, but also to allow for an understanding of the model sensitivities and best ways to apply the policy (Ferguson *et al.*, 2003).

As well as providing insights into national epidemics, models can also be used to estimate the reach and speed of a pandemic, by using network models based on international airline data. The movement of pathogens is helped by the development of transport, allowing for their rapid and effective global spread. Factors such as the increasing susceptible population and the ability of pathogens to cross the species barrier increase the risk of pathogens emerging (Arnal *et al.*, 2011). As will be shown in **Chapter 3**, the global airline network (sometimes referred to as the World Airport Network, or WAN) is growing annually, connecting geographically distant locations. These circumstances allow pathogens to be transmitted across increasingly large geographical ranges through human travel (Tatem and Hay, 2007). **Chapter 6** of this thesis aims to provide an understanding of how the potential origin from which a pandemic may start might affect its international spread.

Although very powerful tools, it must be remembered that mathematical models are never fully accurate. Additional challenges modellers may not be able to take into consideration include anything that cannot be observed or hasn't been measured, such as pathogen evolution (in response to selection pressure and control measures); reporting accuracy (under-reporting may significantly impact the data available for the model and thereby influence the model results); non-homogeneous contact patterns (parts of the population come into more or less contact with other parts of the population); and pathogen ecology (understanding dynamics between multiple hosts and pathogens) (Grassly and Fraser, 2008).

What previous models have shown

Airline passenger travel patterns are known to be good indicators of human international movements and their pathogens (Hosseini *et al.*, 2010). Global events, such as international sporting events and pilgrimages, cause important health threats for the population of the host country by gathering large numbers of travellers from around the world. Consequently, participant population demographics (age, gender, country of origin) are also important to include in any modelling study (Alshammari and Mikler, 2016).

Adequate data relating to instances of the international movements of human infectious disease via airline travel is scarce; making a detailed understanding based solely on notifications of cases related to those instances is difficult, given that in such instances case numbers are likely to be underestimated. Therefore, mathematical models may provide useful tools in better understanding risks posed by international airline travel (Lopez *et al.*, 2016).

The airline network is influenced by the global economy, politics and geography (Guimera *et al.*, 2005), and most seasonal changes in passenger movements occur in the Northern hemisphere, as the largest population resides there (Mao *et al.*, 2015). However, less economically advanced countries are increasingly contributing to the number of outbound tourists as their income and leisure time increase. These countries are also becoming attractive tourist destinations, as tourists now want to visit new and unusual destinations (Wandelt and Sun, 2015). Central cities to the network are not always the most connected to other cities (defined here as a node through which most geodesic paths go to connect to other nodes (Newman, 2003)) in the continent. For example, Atlanta (USA) and Istanbul (Turkey) are very well connected cities, but are not central to the network (not many geodesic paths, or shortest paths, in the network go through them), whereas Anchorage (USA) and Singapore are central to the network but do not feature among the 25 most connected cities (Guimera *et al.*, 2005).

Lopez *et al.* (2016) calculated the risk of travellers importing or exporting diseases by using the force of infection faced by travellers and residents in an endemic country, whereas previous papers have not always considered this risk of infection in the visited country together with airline data to estimate the risk of importation from an endemic to non-endemic country. Lopez *et al.* (2016) use the force of infection but with arbitrary values of airline passengers (of 1,000 passengers), which may be grossly underestimated depending on the country, and thereby give misleading imported infection risks. Therefore, an ideal scenario would be to combine both accurate passenger numbers and the endemic force of infection in the visited country (personal observation).

Aims and objectives of the thesis

Mathematical models can play an important role in providing insights into the predicted spread of an outbreak and may be insightful regarding the most effective control measures to be used to halt the development of a pandemic (Klepac *et al.*, 2018).

The aims of this thesis were to understand what the airline data represents and determine whether the use of airline data were appropriate to understand the international spread of human infectious diseases. Four main objectives were developed to meet these aims, namely:

1) To investigate the data sets and types used by previous mathematical models investigating the international spread of human infectious diseases by airline travel. A checklist to improve the reporting of second-hand data, to be made publicly available, was a sub-objective.
2) To fully understand and compare a detailed airline data set against four independent yet comparable data sets to determine any trends and biases as well as gaining an understanding of the airline network.
3) To model the risk of contracting chikungunya or dengue virus infections faced by travellers from the UK when abroad, compared to the local populations, using a subset of the airline data.
4) Finally, to use a previously generated connectivity matrix derived from the airline data alongside a measure of national health indicators to understand which countries may tend to pose a higher risk for the spread of the next pandemic.

Chapter 2 – The use and reporting of airline passenger data for infectious disease modelling: a systematic review.

Preamble

To understand what data mathematical modellers use to model the international spread of human infectious diseases through airlines, a systematic review was undertaken. As well as looking at the data sources and data types used (bookings or passenger numbers, direct or indirect flights for example), the level of source reporting (allowing independent replication of work) and whether the data were validated were also assessed.

To the best of the author's knowledge, such a review had not been undertaken before, but addressed clear issues in the field. Furthermore, this review allowed for a deeper understanding of the variety of data sources currently available to model human airline movements and their drawbacks. From the results of this review, undertaking a full data description and validation was deemed necessary, and are presented in chapters three and four.

Abstract

A variety of airline passenger data sources are used for modelling the international spread of infectious diseases, with questions existing regarding the suitability and validity of sources used. A systematic review was conducted to identify the sources of airline passenger data used for these purposes, and to assess validation of the data and reproducibility of methodology.

Articles matching the search and inclusion criteria were identified from three databases. From the final 136 articles selected, information regarding the type and source of airline passenger data used was collated, before assessing the studies' reproducibility.

The majority of studies (n=96) used data sources primarily used by the airline industry. Government published data sources were used in 30 studies, and data published by individual airports were used in four studies. Validation of passenger data was conducted in only seven studies. No study was found to be fully reproducible, though eight were partially reproducible.

The author recommends that more effort be made to assess the validity and biases of airline passenger data used for modelling studies, particularly when model outputs are to inform national and international public health policies. Improving reporting standards and more detailed studies to better understand the different biases in different commercial and open access data to permit greater understanding around reproducibility, is also recommended.

Introduction

The international movement of individuals through commercial airline travel has been implicated in the transnational dissemination of many infectious diseases and is thought to be the principle mode of human pathogen transfer between continents. Examples include the global dissemination of the Severe Acute Respiratory Syndrome (SARS) outbreak in 2003, which quickly spread to North America from Hong Kong (Wilder-Smith, 2006). The 2009 influenza pandemic (Fraser *et al.*, 2009), which emerged in Mexico and affected over 208 countries, also saw a similar international dissemination (Al Hajjar and McIntosh, 2010). There is, year-on-year, an increasing number of airline travellers, with a total of 1,186 million international tourist arrivals globally in 2015, a 4.6% increase from 2014 and an additional 510 million arrivals compared to 2000 (Glaesser *et al.*, 2017). Additionally, tourism arrivals from emerging economies are now comparable to those of advanced economy countries, with nations like Mexico and Thailand entering the top 15 of the most visited destinations. This trend in international arrivals is expected to keep rising and reach 1.8 billion arrivals in 2030 (Glaesser *et al.*, 2017). Lower fares and greater availability mean that geographically distant countries and cities are becoming easier and quicker to reach for a greater number of individuals (Saker *et al.*, 2004). Such rapid population movements pose an increasing threat to global populations (Johansson *et al.*, 2011).

The increasing volume of airline passengers seen each year (World Tourism Organisation, 2016) highlights the importance of gaining a better understanding of the dynamics of the airline network and its role in disease spread and control (Mao *et al.*, 2015). There is also a need for accurate prediction of international transmission through passenger flow. Mathematical models are useful tools that can provide an estimated risk of infectious disease importation and exportation by international airline passengers (Lopez *et al.*, 2016), especially in the early stages of an outbreak when accurate reporting may be difficult (Quam and Wilder-Smith, 2016). Models such as the one developed by Lopez *et al.* (2016) use the force of infection in the visited country to determine the risk posed to international passengers, taking an arbitrary number of airline passengers. However, this risk can also extend to new areas when returning passengers carry pathogens back to their residing country, as was the case in Italy in 2007, when a chikungunya outbreak was identified (Quam *et al.*, 2015). Mathematical models of pathogen importation/exportation risks usually entail a function of the infection level in the visited country and the airline passenger volume between the two desired geographical locations, as described in Quam and Wilder-Smith

(2015). Access to accurate and appropriate data sets describing passenger flow between locations is crucial in developing transmission models of global spread (Huang *et al.*, 2013), with which to understand the potential role the airline network may play in the spread of disease, but also to predict future spread, particularly when new threats emerge. However, a variety of data sources have been used (Mao *et al.*, 2015) leading to inconsistency and incomparability between modelling studies. The sources themselves are generally not designed for epidemic modelling purposes. They include data for use within the aviation industry, which may be expensive to access, impose user restrictions, including prohibition from sharing with a third party (Mao *et al.*, 2015; Huang *et al.*, 2013). Open access data sources do exist, but may be geographically restricted, provide information in forms not easily convertible into passenger numbers, or limited in temporal resolution (Mao *et al.*, 2015). Although different sources of passenger data are available, many have drawbacks and inconveniences. Data to model airline passenger movements is, therefore, not necessarily easy to access or appropriate for epidemic modelling purposes. As Balcan *et al.* (2009) state: "The main difficulty in defining a commuting network worldwide is the lack of a global database as opposed to the case of the air-traffic flow".

To gain an overview of the range of airline passenger data sources used by modelling studies, a systematic literature review was designed and conducted. The principal aim of the review was to determine the data types (e.g. passenger numbers, seat capacity) and sources used for the purposes of modelling international infectious disease importation. A secondary aim of the review was to assess the reproducibility of those studies regarding sourcing and use of airline passenger data.

Methods

Search strategy

The search of the literature was conducted on the 2nd of October 2017 using PubMed, Web of Science and Scopus with no restriction on the earliest date of the articles returned. A combination of three sets of search terms was used in this review (#1 AND #2 AND #3). The first set (#1) was: “air” OR “airline” OR “aviation” OR “flight” OR “airport” OR “passenger” OR “transport*” OR “travel*” AND NOT “pollution”. The second set (#2) was: “epidemic” OR “pandemic”. The final set (#3) was: “global” OR “international”. The term ‘Pollution’ was classed as an exclusionary term as initial scoping suggested a large proportion of results included pollution studies, which were deemed irrelevant to this review.

Articles were included if they matched the following inclusion criteria: (1) they were primary and peer-reviewed research; (2) they modelled the international spread of human infectious diseases between at least two countries; (3) the model was parameterised with airline passenger data. We included modelling studies which considered either dynamic models of the transmission process or non-dynamic modelling of infected individual movement. The inclusion of any additional articles, if they were identified as the source of passenger data used within selected articles, and met the three inclusion criteria above was also permitted. Although no language restriction was applied to the searches, articles in a non-English language were excluded during the abstract review if no translated version could be found. Review articles were also excluded, unless specifically addressing the use of airline passenger data. Finally, records which could not be accessed through the University of Liverpool or Lancaster University library records were also excluded.

Following deduplication, the full list of abstracts and titles was first reviewed and included or excluded by at least two reviewers independently. Any disagreement regarding inclusion of an article in the review was then discussed between all reviewers. From the relevant articles selected, the bibliography of each article was searched to find additional relevant articles, based on title and full text. From the final list of selected articles, the full text was accessed and screened for relevance in more detail.

Data collection strategy

From the final selection of articles, information regarding the airline passenger data used in each article was extracted (**Table 2.1**). This information focused on the source, type and validity of data used in the study (**Table 2.1A**), and the reproducibility of data usage judged by pre-defined criteria (**Table 2.1B**). For the purposes of this review, data validation was defined as the comparison of a primary data set used in an article against at least one independent and appropriately comparable set of data. An article was deemed to have validated their data source if they cited another independent and comparable data set and conducted a comparison. To determine their reproducibility, each article was assessed on their reporting of data source using the checklist shown in **Table 2.1B**, and the appropriate score given accordingly. We did not plan or conduct any bias analysis of the selected publication.

Table 2.1: Description of the fields recorded during the literature analysis (Part A) and the reproducibility criteria used to determine reproducibility of articles and sources (Part B).

A. Data description			
Field		*Description*	*Variable*
Article information			
	Authors	List at least first three author names, as on article	*Text*
	Year of publication	Give year of publication	*Date*
	Title	Title of article	*Text*
	Publication name	Name of publication in which the article was published.	*Text*
Data source			
	Commercial data	Commercial databases collecting information about flight routings, aircraft size, number of bookings or passengers. *E.g.*, IATA, OAG, Diio.	Yes/No
	Tourism surveys	Any surveys done in context of tourism. *E.g.*, UNWTO.	Yes/No
	National passenger surveys	Surveys conducted at airports. *E.g.,* International Passenger Survey.	Yes/No
	Airport published information	Data collected and published by airports, may be groups of airports.	Yes/No
	Government immigration data	Data collected by governments on immigration numbers, inbound and outbound.	Yes/No
	Other	*E.g.*, airline published information.	Yes/No
	Unreported or unclear		Yes/No
Data type			
	Seat capacity	Number of seats available on routing.	Yes/No
	Itinerary	Data includes connections, not just origin-destination information.	Yes/No
	Flight numbers	Number of flights between cities/airports/countries or routings.	Yes/No
	Passenger numbers	Data explicitly describes number of passengers travelling.	Yes/No
	Tickets sold	Number of tickets sold or booked per routing.	Yes/No
	Origin-destination information	Data includes origin airport/city/country and destination airport/city/country.	Yes/No
	Direct flight information only	Data does not inform on number of passenger taking connecting flights.	Yes/No
	Unreported or unclear	Insufficient information reported to determine data type.	Yes/No
Data time period			
	Date range of data reported	Period for which the data pertains is reported.	Yes/No
	Date range	State range.	*Text*
Reporting quality (see part B)			
	Fully reproducible	All handling and manipulation of the data is described in detail adequate to enable reproducibility. (reproducibility score = 4)	Yes/No
	Partially reproducible	Important information on handling of the data is missing, or methodology vague or unclear. (reproducibility score = 3)	Yes/No
	Not reproducible	No information on methods and/or data source given and methodology vague or unclear. (reproducibility score $\leq$ 2)	Yes/No
Data validation			
	Is there evidence of data validation?	A comparison was made with an independent and appropriate source of information.	Yes/No
Data usage			
	Transmission model	Airline passenger information is used to parameterise a model of transmission.	Yes/No
	Network analysis	Airline passenger information is described using social network methodology.	Yes/No
	Descriptive or illustrative	Airline passenger information is used to illustrate a transmission risk, but no formal analysis or modelling is performed.	Yes/No
	Other	None of the above (specify or describe what was done).	Yes/No
	Unclear or unreported	Insufficient information to determine data usage.	Yes/No
Pathogen modelled			
	Non-specific	Generic model	Yes/No
	MERS-Corona Virus		Yes/No
	Seasonal influenza		Yes/No
	Pandemic influenza		Yes/No
	Other (specify)		*Text*

(Table 2.1 continues on next page)

(Table 2.1 continued)

B. Reproducibility* Field	Description	Variable
Data accessibility		
Open source	Publicly available, no restrictions on use, no access fees, and source (where online) still accessible as of January 2017.	Yes/No (Yes = +1)
Closed source	Publicly available but restricted access, access may be granted following registration and/or fee; for example, proprietary data.	Yes/No
Not publicly available	Private data, access at discretion of custodian, for example, airport or airline company information.	Yes/No
Reporting clarity of data source		(All Yes = + 1) †
Source identified	The source of the original data is clearly stated.	Yes/No
Dataset named	The specific name of the data set or database from the source is reported.	Yes/No
Access date specified	The date(s) on which data was access is reported	Yes/No
Data type reported	The type or unit represented by the data is reported. *E.g.*, number of flights, number of seat, number of passengers.	Yes/No
Reporting clarity of data usage		
Data handling reported	Data manipulation prior to analysis, including data cleaning and/or aggregation, is reported.	Yes/No (Yes = +1)
Date range of data used		
Data time range reported	The time period of the data is reported.	Yes/No (Yes = +1)
Total reproducibility score	Maximum score = 4	Average total score used if several sources used

* Where studies use a third party's travel model, and if they do not describe the model fully but provide a link or citation, we assessed the cited external documentation cited for reproducibility.
† Authors must receive a 'yes' for all sub-variables for this variable to contribute +1 point to the reproducibility score

Results

From the 4,012 articles identified in the search, 1,465 were identified as duplicates and rejected, resulting in 2,547 articles which went forward for title and abstract screening (**Figure 2.1**). A further 1,130 were rejected at this stage as they did not meet the inclusion criteria. A total of 335 articles were selected based on their title and abstract and read in full. From these, 223 were rejected with the majority (n=87) containing no airline data, 73 were deemed not relevant (did not meet at least two required criteria, such as airline data and model…) and 20 used no model. An additional 19 were country specific, 17 were inaccessible (access to journal or language barrier), five were reviews and two were not focused on human disease movement. After reading the articles in full, 112 were selected as relevant to this review. Finally, 24 additional articles, not detected by the search but identified through reading the bibliography of accepted articles were included after being read in full to determine their relevance.

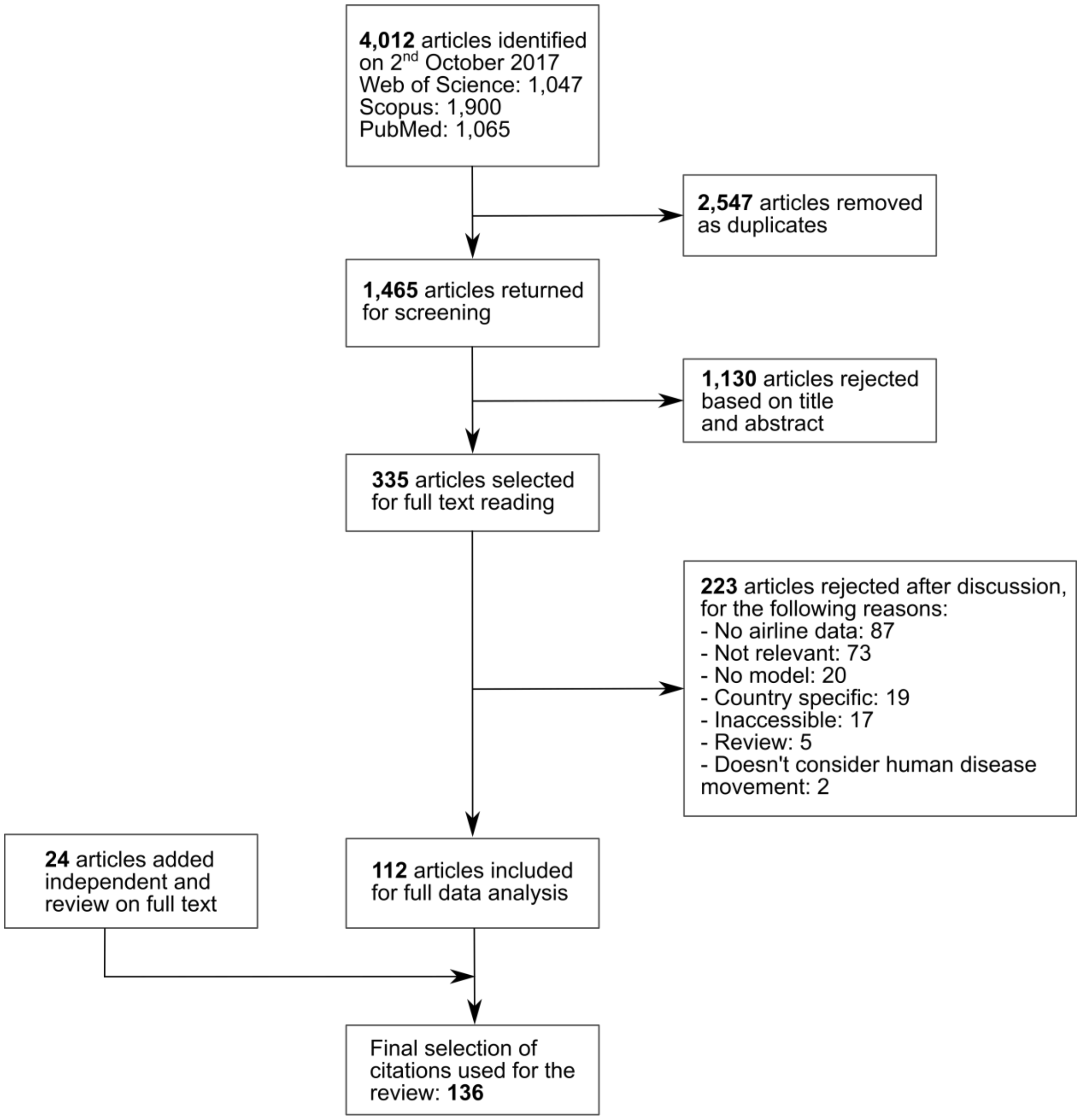


Figure 2.1: Flowchart of the article selection undertaken during the systematic literature review.

The year of publication for the 136 articles ranged from 1985 to 2017, with the largest number of articles (n=17) published in 2016 (a detailed list of accepted titles can be found in **Table 2.5**). In the twenty years following from the Rvachev and Longini (1985) publication, a total of seven articles were published [41; 43; 58; 66; 83; 84; 115].

A wide range of data sources have been used to model passenger flow between countries with a total 45 distinct sources identified here (**Table 2.2**). Commercial or industry data sources were most often used (14 sources, used in 133 articles), followed by governmental data (14 sources, used in 32 articles). Of the commercial data sources, those most often acknowledged were International Air Transport Association (IATA) (62 articles) and OAG (38 articles); where a database was named from these sources, OAG MAX was the most frequently used (3 articles) followed by t100 (2 articles) and Traffic Analysis Tool (1 article). A range of other industry-orientated data sources were cited, including Diio (airline market information), Amadeus (travel reservations database), Feeyo (Chinese-based flight scheduler), and an open-access publicly-contributed database (OpenFlights.org). Four articles used passenger surveys, such as TravelPac from the United Kingdom's (UK) Office for National Statistics (ONS), and eight articles used tourism surveys (**Table 2.2**). Five articles used information published by airports, and four other sources were reported (the social media site Twitter, two aircraft manufacturers and EuroStat).

Most data sources used described origin-destination information (n=91, 67%) or passenger numbers (n=73, 54%) (**Table 2.3**). Data describing direct flights only were used more often than data describing full passenger itineraries: n=33 and n=27, respectively. Of those using IATA as a data source, 15 used direct flights only (1; 6-8; 13; 29-31; 33-35; 64; 105; 109; 119) and of those using OAG, 11 used direct flights only (7; 8; 10; 18; 22; 45; 82; 87; 119; 123; 124). Finally, eight articles indirectly used IATA data by using the online modelling tool GLEAMviz (1; 5; 6; 8; 56; 99; 103; 127) (2017) and two by using BioDisapora (now Bluedot.global) (76; 109) (Bluedot.global).

Table 2.2: Data sources identified in the selected articles, grouped by sector.

Data Source	Number of articles citing*	Article reference number (see Table 2)
Commercial / industry	**133 (62%)**	
International Air Transport Association (IATA)	62	
Unspecified	58	1; 4-8; 11-13; 19; 23; 28-36; 38; 43; 44; 47; 51-53; 55; 56; 58; 63; 64; 66; 72-75; 77-79; 89; 90; 99; 102-105; 109; 111; 113; 114; 118-120; 127; 132; 133; 135
Air Passenger Market Analysis	1	48
Airport Intelligence Services – Passenger data	1	134
International Travel Statistics	1	62
Passenger Intelligence Services	1	46
OAG	38	
Unspecified	32	7; 8; 10; 18; 19; 22; 38; 42; 44; 45; 55; 56; 58; 66; 72; 82-84; 87; 88; 93; 94; 99; 104; 114-116; 119; 122; 124; 125; 127; 135
OAG MAX	3	14; 54; 123
t 100	2	68; 69
Traffic analyser	1	112
International Civil Aviation Organization (ICAO)	12	
Unspecified	11	25; 38; 44; 57; 72; 83; 84; 115; 120; 126; 129
Traffic by Flight Stage	1	58
Air Transport Statistics	3	38; 84; 115
Airports Council International	1	78
Amadeus	1	65
Back Aviation Solutions Incorporated	4	38; 44; 58; 72
CapStat	1	98
Diio	3	59; 67; 70
Feeyo	1	24
Landings.com	1	65
OpenFlights.org	4	37; 81; 96; 97
OurAirports.com	1	37
Turism.se (Swedish Tourist and Travel Database)	1	41
Tourism surveys	**8 (4 %)**	
Icelandic Tourist Board	1	130
Singapore Tourism Board	1	121
UNWTO	5	40; 61; 101; 131; 136
US Office of Travel and Tourism Industries	1	124
National Passenger surveys	**4 (2 %)**	
Brazilian Ministry of Tourism	1	40
UK Office for National Statistics	3	61; 101; 107
Airport published information	**12 (7 %)**	
Amsterdam Airport (Schipol)	1	3
Beijing Capital International Airport	2	62; 63
German airports (Hannover, Frankfurt, Hamburg, Munich)	1	3
Hong Kong International Airport	2	62; 63
Keflavik Airport	1	130
London airports (Heathrow, Gatwick, Stansted, Luton)	1	3
Los Angeles International Airport	1	3
Madeira Airport	1	85
Teheran Airport	1	3
Venice Airport	1	3

(Table 2.2 continues on next page)

(Table 2.2 continued)

Government published information	**32 (15%)**	
US Department of Transport	14	15; 17; 37; 38; 44; 49; 50; 58; 68; 69; 72; 86; 100; 128
Australian Department of Transport	2	83; 84
Australian International Airport Traffic	4	38; 84; 84; 115
Brazilian Ministry of Tourism	1	92
Hong Kong Tourism Board	1	27
Japan National Tourism Organization	1	110
Malaysian Department of Statistics	1	2
Mexican Secretary Communication and Transport	1	60
National Statistics China	1	57
Saudi General Authority of Civil Aviation	1	78
Singapore Tourism sector performance	1	91
Statistics Canada	1	80
Statistics Iceland	1	130
UK Civil Aviation Authorities	2	9; 20
Other sources	**11 (5 %)**	
Airbus Industries	3	38; 44; 58
Boeing Corporation	3	44; 38; 58
EuroStat	4	50; 95; 106; 120
Twitter	1	16
Unclear or unreported†	**13 (6 %)**	
Unclear or unreported	13	21; 26; 38; 39; 44; 58; 71; 76; 80; 84; 93; 115; 117

* Some articles used more than one data source.

Table 2.3: Frequency of use of each data type identified within the literature review.

Data type*	**Number of articles citing** (% of selected articles)	**Article reference number** (see Table 2)
Includes origin-destination information	91 (67%)	3; 5-8; 10-14; 19; 20; 22; 23; 25; 26; 29; 32; 34; 35-37; 40-43; 45-51; 55; 57-61; 65-68; 70; 73-76; 78; 79; 81; 82; 87-90; 93; 96-100; 102; 104; 106-114; 118-120; 122-136
Passenger numbers	73 (54%)	3; 9; 12; 13; 15; 17; 20; 21-23; 25; 26; 33; 35; 37; 39-41; 43; 45-51; 57-63; 66; 70; 73-76; 78; 86; 89-92; 98; 100; 101-104; 106; 108-114; 118-122; 124-126; 128-134; 136
Direct flights only	33 (24%)	1; 3; 6-8; 10; 13; 18; 22; 25; 29-31; 33-35; 37; 45; 49; 57; 64; 67; 70; 81; 82; 87; 105; 106; 109; 119; 120; 123; 124
Full itinerary	27 (20%)	11; 13; 14; 23; 46-48; 51; 68; 73-75; 77; 79; 89; 96; 104; 109; 113; 114; 118-120; 125; 133-135
Unreported or unclear	25 (19%)	2; 4; 16; 27; 37; 38; 44; 52; 53; 56; 58; 63; 69; 71; 72; 80; 83-85; 92; 93; 95; 107; 115; 117
Seat capacity	24 (18%)	1; 6-8; 10; 18; 28-31; 35; 36; 55; 64; 67; 81; 82; 88; 99; 116; 123; 124; 127; 135
Flight numbers	13 (10%)	14; 17-19; 42; 54; 60; 65; 86; 87; 93; 94; 97
Tickets sold	3 (2%)	11; 24; 37

* An article may use multiple data types

From the measures of methodological reproducibility described in **Table 2.1B**, it became apparent that of the 45 total sources identified, 26 (58%) were open source, and 11 (24%) were closed source. The date range of the data (start and end date) was reported in 58% (n=79) of the studies, and an access date (date of data download) was stated in 25% (n=34) of the sources used. Data validation as previously defined was performed in 5% (n=7) of the articles collected (16; 41; 53; 75; 119; 120; 127). Only 40 articles (29%) reported performing any data cleaning or manipulation before using their data set. Given the set of standards established to determine an article's reproducibility, no article was considered fully reproducible; eight (6%) were deemed partially reproducible (score of 3 or above), where some information regarding the description and use of passenger data were reported (9; 16; 24; 41; 47; 49; 110; 130).

The majority of articles (n=115, 85%) modelled the global spread of infectious diseases, while the analysis of the airline network itself was the next most common purpose (n=11, 8%). Five articles used passenger data for descriptive or illustrative purposes (28; 33; 34; 41; 77), two articles used the data for passenger screening simulations (17; 86) and two articles described a public health tool development (7; 37). Of the pathogens modelled, pandemic influenza was the most frequent subject of the models (n=40, 29%) (**Table 2.4**). Generic models not focussing on a specific pathogen were also common (n=23, 17%).

Table 2.4: Pathogens modelled within selected articles.

Pathogen name*	**Number of articles citing** (% of selected articles)	**Article reference number** (see Table 2)
Generic model (no specific pathogen)	23 (17%)	1; 4; 6; 10; 18; 30; 31; 34; 35; 38; 52-54; 63; 69; 76; 77; 94; 106; 107; 117; 123; 129
Influenza virus		
Pandemic	40 (29%)	3; 5; 7; 8; 17; 21; 24; 26; 27; 29; 36; 42-45; 49; 58; 60; 62; 64-67; 72; 73; 75; 78; 80-84; 86-88; 95; 115; 127; 128; 136
Seasonal	7 (5%)	9; 15; 16; 71; 93; 100; 130
MERS-Corona Virus	7 (5%)	22; 48; 79; 97; 102; 103; 105
Other		
Chikungunya virus	6 (4%)	23; 70; 74; 98; 120; 124
Vibrio cholerae	1 (0%)	37
Clostridium difficile	1 (0%)	28
Dengue virus	17 (13%)	25; 47; 50; 51; 85; 91; 92; 98; 108-110; 118-121; 126; 133
Ebola virus	7 (5%)	13; 55; 99; 104; 112; 116; 135
Hepatitis A virus	1 (0%)	15
Human Immunodeficiency Virus (HIV)	1 (0%)	2
Japanese Encephalitis virus	1 (0%)	61
Plasmodium sp. (malaria)	5 (4%)	15; 39; 101; 124; 125
Measles virus	1 (0%)	134
Polio virus	1 (0%)	132
Severe Acquired Respiratory Syndrome (SARS)	6 (4%)	14; 19; 32; 33; 57; 114
Smallpox virus	1 (0%)	56
Salmonella typhi and *enterica*	1 (0%)	41
Vector importation	1 (0%)	122
West Nile virus	1 (0%)	20
Yellow fever virus	3 (2%)	40; 68; 131
Zika virus	9 (7%)	11; 12; 46; 59; 89; 90; 96; 111; 113

* An article may use multiple pathogens

Discussion

The purpose of this review was to assess the source and usage of airline passenger data used in mathematical models of international infectious disease spread. A total of 136 articles were identified as meeting the inclusion criteria, from which a total of 45 unique data sources were identified.

A variety of sources were identified in these articles, with the majority of them produced by and for the commercial aviation industry. Examples of this type of data source include the International Air Transport Association (IATA), OAG and the International Civil Aviation Organization (ICAO). These commercial sources provide information from the aviation industry for use within that industry, and are marketed as being detailed and accurate. The data resolution can be high: for example, passenger data is available stratified by routing (including stopovers), fare class, point of origin, and time period. There are often user restrictions on the use of the data, and financial charges made for access (Mao *et al.*, 2015). This type of data can be deemed closed data, meaning it is publicly available but at a price and with restricted access. Furthermore, the methodology underpinning data collection is generally undisclosed, and as such it is difficult for researchers to assess the quality, representability and biases of the data. Although these data sources may have a number of subsets representing different data types, a more accurate reporting of the data sets, including name of subsets used and date of access, among other criteria, are not often reported by authors.

A number of data sources identified in the review are open-access and include passenger data published by individual airports, data compiled and released by government agencies (for example, UK Office for National Statistics), and information derived from tourism surveys. Although freely available to access, these data sets may not provide the resolution of information required by modelling studies, as they typically are limited to passengers departing from or arriving at a specific geographical region, or are aggregated over long-time periods (annual or quarterly data). Additionally, the collection methodology is not always reported for such data sources. Combining information from such sources represents a considerable data challenge.

International travel data describing direct flights only were used more often than those with full itinerary information. Data based on direct flights excludes information on connecting passengers, and will therefore underestimate the number of passengers travelling to a specific destination. This limitation is likely to introduce bias, underestimating passenger flow between distant or poorly served locations, and overestimating passengers travelling shorter distances (Johansson *et al.*, 2011). This bias has implications for public health planning, as some locations or countries may have an apparent lower risk of importation events due to the lack of direct flights from source countries known to have many infection events. This may explain the discrepancy between studies during the West Africa Ebola epidemic of 2014-15 where several studies suggested the USA was at relatively low risk of importation due to the suspension of direct flights. The USA did however receive two importations through air travel from the affected area, one due to a passenger reaching their final destination through indirect flights and the second from a returning healthcare worker (Bogoch *et al.*, 2015; Gomes *et al.*, 2014; European Centre for Disease Prevention and Control, 2015).

When considering international travel patterns for public health purposes accessing information on the number of passengers travelling from an origin to a destination is the most relevant. However, we found several articles used data for which the unit of measurement was not the number of passengers. Several data sources used describe passenger traffic in terms of seat capacity – literally the number of seats on aircraft flying between two specific airports – for which assumptions must be made regarding how full individual flights are and how this may or may not vary according to seasons. Additionally, this data type cannot take into account the full routing of a passenger, which must therefore be inferred from the data, or state that only direct flights are considered for the study in question. The variety of data types used for epidemic modelling purposes perhaps reflects the lack of a widely accepted and accessible data source, and this variation in data unit could lead to differences in conclusion between modelling studies.

To ensure reproducibility by others, information regarding the source and type of the data used, the date of access, and any cleaning or manipulation conducted prior to use should be reported. This analysis showed this standard is rarely attained. Reporting the date of access is important as several data providing companies update their data monthly, with retrospective adjustments of values (OAG, 2015). Few studies (n=34, 25%) reported the date of access to the data set. Acknowledging any data cleaning/manipulation is also important for reproducibility (Yale Law School Roundtable on Data and Code Sharing, 2010): for example, if the authors are considering passengers departing or arriving from cities rather than airports, but the data were collected at the airport level, the aggregation of passenger numbers from each airport to the city should be acknowledged by the authors. For additional clarity, it would be useful for authors to report the stage at which the data were aggregated to city level: whether this was part of the original data, or if this was a data manipulation done by the authors. Additionally, at the time of writing there is a limited understanding of the sensitivity of this level of data (city level) and how it compares to airport level data and other aggregated data sets, requiring further analytical work. Overall, the majority of articles were deemed to have methods that were unreproducible, and while eight studies were deemed partially reproducible none were considered to be fully reproducible. As it is the author's responsibility to ensure accurate reporting for all aspects of their methodology, the findings of this review suggest that authors of international disease modelling studies should aim to improve their reporting of airline passenger data source and usage. Authors are advised to reference the fields reported in **Table 2.1B**, at a minimum, when using any data sets.

Data validation is often required to ensure that the data collected is fit for purpose, free from biases, and is an accurate reflection of the subject or process being described. Validation of airline passenger data is particularly important to conduct if the passenger data is sourced from a commercial company with limited or no collection methodology disclosed. Only seven articles reported validation with at least one independent or appropriately comparable set of observations. While there is no acknowledged 'gold standard' data set, governmental open source data, such as that from the US Department of Transport or Office for National Statistics, do at least have published methodology on which potential biases may be identified.

Human travel introduces pathogens to susceptible populations or with little awareness, allowing for potential further spread and rising incidence. In the articles considering a pathogen, the majority used viral transmission or importation. Only three articles were focused on bacteria (*Vibrio cholerae*, *Clostridium difficile* and *Salmonella enterica serotypes typhi* and *paratyphi*), despite the known importance of international travel for their capacity to initiate epidemics following importation, Haiti cholera outbreak in 2010 for example, (Chin *et al.*, 2011) and the global dissemination of antibacterial resistance (Amesh *et al.*, 2018; Bernasconi *et al.*, 2016; Holmes *et al.*, 2016; Lepelletier *et al.*, 2011; Okeke and Edelman, 2001). Pandemic influenza was the virus most often considered by the reviewed articles, which perhaps reflects the global significance of pandemic events and the ease with which pandemic strains have spread historically. The other non-influenza viruses noted in these studies have all initiated outbreaks following introduction through international travel, namely MERS-Corona virus in South Korea (Cowling, 2015), dengue virus in the Portuguese islands of Madeira (off the coast of Western Africa) (Semenza *et al.*, 2014) and chikungunya virus in the Caribbean (leading to imported cases in the United States) (Khan *et al.*, 2014) and Italy (Rezza *et al.*, 2007). Finally, the accurate modelling of importation risks for specific pathogens may require very high resolution passenger data, particularly where routings are indirect and the total travel time from origin to destination is important in the screening efficacy due to incubation periods (Read *et al.*, 2015).

To the best of the author's knowledge, direct comparisons of commercial and open access data sets, or commercial data sets between themselves, have not yet been accomplished, preventing an informed decision on which data sets are more suitable to represent airline passengers. Although a direct comparison between commercial data sets is likely to be very informative for the modelling community, it is likely to be very expensive. Additionally, the presence of a single data-set that is agreed by the community to best represent international (and national) airline passenger flow would be ideal, though may be difficult to realise given proprietorial restrictions of certain data sets. None-the-less, work is being done regarding epidemiological data to gather infectious disease outbreak into a central and unified database structure (Finnie *et al.*, 2016), and the field should aspire to work collaboratively with industrial data providers to realise accurate passenger data available for research, particularly during global public health emergencies.

Strengths and limitations

The screening and selection of articles was done in a systematic manner and by two independent reviewers to ensure all relevant articles downloaded were included in the selection of articles to be read in full. The full reference list of accepted articles was read to find additional relevant articles. Although a number of articles were found when reading the reference lists, the author is confident that this selection is a good representation of the range of airline data used. Additionally, no other review that the author is aware of was focused on the analysis of the validity and reproducibility of the data used for mathematical models.

Limitations of this study include not contacting authors regarding their methods. Additionally, by limiting the articles to international spread only, some articles which focused primarily on within country spread, such as Bozick and Real (2015), Charu *et al.* (2017) and Epstein *et al.* (2007), among others, were deliberately left out, even though they may use relevant data sources.

Table 2.5: Reference list of the articles analysed (n=136) for the literature review, detailing the name of their data sources, whether the data was validated and the article's reproducibility score.

Num-ber	Authors	Year	Journal reference	Sources used	Valida-tion	Reprod ucibility score *
1	Ajelli M, B Gonçalves, D Balcan, *et al*	2009	BMC Infect Dis; **10**:190	IATA	No	0
2	Apenteng, O. and Ismail, N.	2014	Transactions on Engineering Technologies; 381-389	Malaysian Department of Statistics	No	2
3	Apolloni A, C Poletto and V Colizza	2013	BMC Infect Dis; **13**:176	Airports: LAX, Teheran, London, Amsterdam, Venice, German	No	(0+0+1+ 0+1+0) 0.33
4	Arino, J. and Khan, K	2014	Analyzing and Modeling Spatial and Temporal Dynamics of Infectious Diseases (book)	IATA	No	1
5	Bajardi P, C Poletto, J Ramasco, *et al*	2011	PLoS ONE; **6**	IATA	No	0
6	Balcan D, V Colizza, B Gonçalves, *et al*	2009	PNAS; **106;** 21484-21489	IATA	No	0
7	Balcan, D. Goncalves, B. Hu, H. et al	2010	J Comput Sci; **1**; 3; 132-145	IATA and OAG	No	(0+0) 0
8	Balcan D, H Hu, B Goncalves, *et al*	2009	BMC Medicine; **7**; 49	IATA and OAG	No	(0+0) 0
9	Bedford, T. Riley, S. Barr, I. G. *et al*	2015	Nature; **523**; 7559; 217-20	Civil Aviation Authority	No	3
10	Bobashev G, R Morris and M Goedecke	2008	PLoS ONE; **3**:9	OAG	No	2
11	Bogoch, I. Brady, O. Kraemer, M. *et al*	2016	Lancet ID; **16**; 11; 1237-1245	IATA	No	2
12	Bogoch, I. Brady, O. Kraemer, M. *et al*	2016	Lancet; **387**; 10016; 335-336	IATA	No	2
13	Bogoch I, M Creatore, M Cetron, *et al*	2015	Lancet; **385**; 29-35	IATA	No	2
14	Bowen Jr J and C Laroe	2006	Geogr J; **172**; 130-144	OAG (OAG MAX)	No	1
15	Brannen, D. Alhammad, A. Branum, M. *et al*	2016	Scientifica; 8258946	US Department of Transportation (Air Carrier Activity Information System)	No	2
16	Brennan S, A Sadilek and H Kautz	2013	IJCAI; 2783-2789	Twitter	No	3
17	Brigantic R, Malone J, Muller G, *et al*	2009	Int J Risk Assess Manag; **12**; 290-310	US Department of Transport	No	1
18	Brockmann D and D Helbing	2013	Science; **342**; 1337-1342	OAG	No	0
19	Brockmann D, L Hufnagel and T Geisel	2007		IATA and OAG	No	(0+0) 0
20	Brown, E. Adkin, A. Fooks, A. *et al*	2012	Vector-Borne Zoonot; **12**; 4; 310-320	Civil Aviation Authorities	No	2
21	Caley P, N Becker and D Philip	2007	PLoS ONE; **2**:1	Unknown	No	0
22	Carias, C. O'Hagan, J. Jewett, A. *et al*	2016	Emerging Infec Diseases; **22**; 4; 723-725	OAG	No	2
23	Cauchemez S, M Ledrans, C Poletto, *et al*	2014	Euro Surveill; **19**; 20854	IATA	No	1
24	Chang C, C Cao, Q Wang, *et al*	2010	Chin. Sci. Bull.; **55**; 3030-3016	Feeyo	No	3
25	Cheng, Q. Jing, Q. Spear, R. *et al*	2017	PLoS Negl Trop Dis; **11**; 6	ICAO	No	1
26	Chong, K. Fong, H. and Zee, C.	2014	Epidemiol Infect; **142**; 5; 955-63	Unknown	No	2
27	Chong, K. and Zee, B.	2012	BMC Infect Dis; **12**	Hong Kong Tourism Board	No	1
28	Clements A, R Magalhães, A Tatem, *et al*	2010	Lancet Infect Dis; **10**; 395-404	IATA	No	0
29	Colizza V, A Barrat, M Barthelemy, *et al*	2007	PLoS Med; **4**; 95-110	IATA	No	0
30	Colizza V, A Barrat, M Barthelemy, *et al*	2006	Bull Math Biol; **68**; 8; 1893-921	IATA	No	1
31	Colizza V, A Barrat, M Barthelemy, *et al*	2006	PNAS; **103**; 7; 2015-20	IATA	No	1

(Table 2.5 continues on next page)

(Table 2.5 continued)

Num-ber	Authors	Year	Journal reference	Sources used	Valida-tion	Reproducibility score *
32	Colizza V, A Barrat, M Barthelemy, *et al*	2007	BMC Med;**5**; 34	IATA	No	0
33	Colizza V, A Barrat, M Barthelemy, *et al*	2008	Biophys Rev Lett; **3**; 1-2; 217-226	IATA	No	0
34	Colizza V, M Barthélemy, A Barrat, *et al*	2007	C R Biol; **330**; 364-374	IATA	No	0
35	Colizza V and A Vespignani	2008	J Theor Biol; **251**; 450-467	IATA	No	0
36	Cooper B, R Pitman, W Edmunds, *et al*	2006	PLoS Med; **3**; 845-855	IATA	No	1
37	Corley C, Lancaster M, Brigantic R, *et al*	2012	Proc ACM SIGSPATIAL Int Conf Adv Inf; 81-86	US Department of Transport; OpenFlights.org; OurAirports.com	No	(2+1+1) 1.33
38	Daniel, W. Hengartner, N. Rivera, M. *et al*	2013	Math Biosci; **242**; 1; 1-8	Grais, *et al* (2003); Rvachev and Longini (1985)	No	(0.4+0.6) 0.47
39	Dembele, B. and Yakubu, A.	2017	Math Biosci Eng; **14**; 1; 95-109	Unknown	No	0
40	Dorigatti, I. Hamlet, A. Aguas, R. *et al*	2017	Euro Surveill ; **22**; 28	UNWTO; Brazilian Ministry of Tourism	No	(2+3) 2.5
41	Ekdahl, K. De Jong, B. and Andersson, Y.	2005	J Travel Med; **12**; 4; 197-204	Swedish Tourist and Travel Database	Yes	3
42	Epstein J, D Goedecke, F Yu, *et al*	2007	PLoS ONE; **2**:5	OAG (OAG MAX)	No	0
43	Flahault A, S Deguen and A Valleron	1994	Eur J Epidemiol; **10**; 471-474	IATA	No	0
44	Flahault A, E Vergu, L Coudeville, *et al*	2006	Vaccine; **24**; 6751-6755	US Department of Transport; OAG; IATA; ICAO; Back Aviation Solutions; Air Transportation Statistics; Australian International Arrivals; Airbus Industries; Boeing corporation; unknown	No	(2+1+1+1+1+0+0+0+1+1) 0.5
45	Fraser, C. Donnelly, C. Cauchemez, S. *et al*	2009	Science ; **324**; 5934; 1557-1561	OAG	No	2
46	Gardner, L. Chen, N. and Sarkar, S.	2017	PLoS Negl Trop Dis; **11**; 3	IATA (Passenger Intelligence Services)	No	2
47	Gardner, L. and Sarkar, S.	2013	PLoS One; **8**; 8	IATA	No	3
48	Gardner, L. Chughtai, A. and MacIntyre, C.	2016	J Travel Med; **23**; 6	IATA (Air passenger market analysis)	No	2
49	Gardner, L. Fajardo, D. and Waller, S.	2012	Transport Res Rec; 2300; 13-21	US Department of Transport	No	3
50	Gardner, L. Fajardo, D. Waller, S. *et al*	2012	J Trop Med;	US Department of Transport; Eurostat	No	(3+2) 2.5
51	Gardner, L. and Sarkar, S.	2015	Transport Res Rec; **2501**; 25-30	IATA	No	2
52	Gautreau A, A Barrat and M Barthélemy	2007	J Stat Mech Theor Exp; **9**	IATA	No	0
53	Gautreau A, A Barrat and M Barthélemy	2008	J Theor Biol; **251**; 509-522	IATA	Yes	0
54	Goedecke M, G Bobashev and F Yu	2007	Proc Winter Simul Conf; 1538-1542	OAG (OAG MAX)	No	2
55	Gomes M, Y Pastore, L Rossi, *et al*	2014	PLoS Curr; **6**	IATA; OAG	No	(0+0) 0
56	Gonçalves, B. Balcan, D. and Vespignani, A.	2013	Sci Rep-UK; **3**	IATA; OAG	No	(0+0) 0
57	Goubar A, Bitar D, Cao W, *et al*	2009	Epidemiol Infect; **137**; 1019-1031	ICAO; National Bureau of Statistics of China	No	(1+1) 1
58	Grais R, J Ellis and G Glass	2003	Eur J Epidemiol; **18**; 1065-1072	US Department of Transport; OAG; IATA; ICAO (Traffic by Flight Stage); Back Aviation Solutions; Air Transportation Statistics; Australian International Arrivals; Airbus Industries; Boeing corporation; unknown	No	(2+0+0+0+0+0+0+0+0+1) 0.4
59	Grills, A. Morrison, S. Nelson, B. *et al*	2016	MMWR; **65**; 28; 711-715	DiiO	No	1
60	Hanvoravongchai P and R Coker	2011	Southeast Asian J Trop Med Public Health; **42**; 1093-99	Mexican Secretary of communication and transport	No	2
61	Hatz C, J Werlein, M Mutsch, *et al*	2009	J Travel Med; **16** ; 3; 200-3	UNWTO; UK Office for National Statistics	No	(1+3) 2

(Table 2.5 continues on next page)

(Table 2.5 continued)

Num-ber	Authors	Year	Journal reference	Sources used	Valida-tion	Reproducibility score *
62	Hollingsworth D, N Ferguson and R Anderson	2006	Nature Med; **12**; 5; 497-499	Beijing Capital International Airport (Traffic Data); Hong Kong International Airport (Provisional Civil International Air Traffic Statistics) ;IATA	No	(1+1+0) 0.67
63	Hollingsworth D, N Ferguson and R Anderson	2007	Emerg Infect Dis; **13**; 1288-1294	IATA (International Travel Statistics); Hong Kong International Airport; Beijing Capital Airport	No	(1+1+0) 0.67
64	Hosseini P, S Sokolow, K Vandegrift, *et al*	2010	PLoS ONE; **5**:9	IATA	No	1
65	Hsu, C. and H. Shih ,	2010	Accid Anal Prev; **42**; 93-100	Amadeus; Landing.com	No	(0+1) 0.5
66	Hufnagel L, D Brockmann and T Geisel	2004	PNAS; **101**; 15124-15129	IATA; OAG	No	(0+0) 0
67	Hwang G, P Mahoney, J James, *et al*	2012	Travel Med Infect Dis; **10**; 32-42	DiiO	No	2
68	Johansson M, N Arana-Vizcarrondo, B Biggerstaff, *et al*	2012	Am J Trop Med Hyg; **86**; 394-358	OAG (Traffic Analyser); US Department of Transport	No	(0+1) 0.5
69	Johansson M, N Arana-Vizcarrondo, B Biggerstaff, *et al*	2011	PLoS ONE; **6**:7	OAG (Traffic Analyser); US Department of Transport	No	(0+1) 0.5
70	Johansson, M. Powers, A. Pesik, N. *et al*	2014	PLoS ONE; **9**; 8	DiiO	No	2
71	Kenah E, D Chao, L Matrajt, *et al*	2011	PLoS ONE; **6**:5	Unknown	No	0
72	Kernéis S, R Grais, P Boëlle, *et al*	2008	PLoS ONE; **3**:1	US Department of Transport; OAG; IATA; ICAO; Back Aviation Solutions	No	(2+0+0+0+0) 0.4
73	Khan K, J Arino, W Hu, *et al*	2009	N Engl J Med; **361**; 212-214	IATA	No	1
74	Khan, K. Bogoch, I. Brownstein, J. *et al*	2014	PLoS Curr; **6**	IATA	No	2
75	Khan K, R Eckhardt, J Brownstein, *et al*	2013	Bull World Health Organ; **91**; 368-376	IATA	Yes	2
76	Khan, K. Freifeld, C. Wang, J. *et al*	2010	CMAJ; **182**; 6; 579-583	Unknown	No	2
77	Khan K, S McNabb, Z Memish, *et al*	2012	Lancet Infect Dis; **12**; 222-230	IATA	No	1
78	Khan K, Z Memish, A Chabbra, *et al*	2010	J Travel Med; **17**; 75-81	ACI; Saudi Arabia Authority of Civil Aviation; IATA (Worldwide passenger ticket sales)	No	(1+2+0) 1
79	Khan K, J Sears, V Hu, *et al*	2013	PLoS Curr; **5**	IATA	No	2
80	Knipl D, G Röst and J Wu	2013	SIAM J Appl Dyn Syst; **12**; 1722-1762	Statistics Canada; unknown	No	(1+1) 1
81	Lawyer, G.	2016	BMC Infect Dis; **16**; 70	OpenFlights.org	No	2
82	Lemey P, A Rambaut, T Bedford, *et al*	2014	PLoS Pathog; **10**:2	OAG	No	1
83	Longini Jr, I. M.	1987	Math Biosci; **90**; 367-383	Rvachev and Longini (1985)	No	0.6
84	Longini I, Fine P. and S. Thacker	1986	Am J Epidemiol; **123**; 383-391	Air Transport Statistics; Australian international airport traffic dynamics; ABC World Airways Guide; OAG; ICAO	No	(0+1+0+0+1) 0.4
85	Lourenço, J. and M. Recker,	2014	PLoS Neglected Trop Dis; **8**:8	Madeira Airports	No	1
86	Malone J, R Brigantic, G Muller, *et al*	2009	Travel Med Infect Dis; **7**; 181-191	US Department of Transport	No	1
87	Marcelino, J. and M. Kaiser,	2009	PLoS Curr; **1**	OAG	No	2
88	Marcelino J. and M. Kaiser	2012	PLoS Curr; **4**	OAG	No	2
89	Massad, E. Burattini, M. Khan, K. *et al*	2017	Epidemiol Infect ; **145** ; 11 2303-2312	IATA	No	1
90	Massad, E. Tan, S-H. Khan, K. *et al*	2016	Global Health Action; **9**; 1	IATA	No	1
91	Massad, E. and A. Wilder-Smith	2009	J Travel Med; **16**; 3; 191-3	Singapore Tourism Sector Performance	No	2
92	Massad, E. Wilder-Smith, A. Ximenes, R. *et al*	2014	Memorias do Instituto Oswaldo Cruz; **109**; 3; 394-7	Brazilian Ministry of Tourism	No	1

(Table 2.5 continues on next page)

(Table 2.5 continued)

Number	Authors	Year	Journal reference	Sources used	Validation	Reproducibility score *
93	Matrajt L, Halloran E and Longini I	2013	PLoS Comput Biol; 9:3	OAG (OAG MAX); unknown	No	(2+0) 1
94	Meloni, S. Perra, N. Arenas, A. *et al*	2011	Sci Rep; **1**; 62	OAG	No	2
95	Merler, S. and M. Ajelli,	2010	Proc Biol Sci; **277**; 1681; 557-65	Eurostat	No	2
96	Nah, K. Mizumoto, K. Miyamatsu, Y. *et al*	2016	PeerJ; **2016**; 4	OpenFlights.org	No	2
97	Nah, K. Otsuki, S. Chowell, G. *et al*	2016	BMC Infect Dis; **16**; 356	OpenFlights.org	No	2
98	Napoli, C. Salcuni, P. Pompa, M. *et al*	2012	J Travel Med; **19**; 5; 294-7	CapStat	No	1
99	Pastore-Piontti, A. Zhang, Q. Gomes, M. *et al*	2016	Mathematical and Statistical Modeling for Emerging and Re-emerging Infectious Diseases; 39-56	IATA;OAG	No	(1+1) 1
100	Paul, M. Held, L. and A. Toschke	2008	Stat Med; **27**; 6250-6267	US Department of Transport	No	2
101	Pinset, A. Read, J. Griffin, J. *et al*	2014	Malaria J; **13**; 298	UNWTO; UK Office for National Statistics	No	(2+1) 1.5
102	Poletto, C. Boelle, P. and V. Colizza,	2016	BMC Infect Dis; **16**; 1; 448	IATA	No	1
103	Poletto, C. Boelle, P. and V. Colizza,	2016	Epidemics; **15**; 1-9	IATA	No	0
104	Poletto, C. Gomes, M. Pastore Y Piontti, A. *et al*	2014	Euro Surveill; **19**:42	IATA; OAG	No	(1+1) 1
105	Poletto, C. Pelat, C. Lévy-Bruhl, D. *et al*	2014	Euro Surveill; **19**:23	IATA	No	0
106	Poletto, C. Tizzoni, M. and V. Colizza,	2012	Sci Rep; **2**; 476	EuroStat	No	1
107	Poletto, C. Tizzoni, M. and V. Colizza,	2013	J Theor Biol; **338**; 41-58	UK Office for National Statistics	No	1
108	Polwiang, S.	2015	J Travel Med; **22**; 3; 194-9	Department of Tourism of Thailand	No	2
109	Quam, M. Khan, K. Sears, J. *et al*	2015	J Travel Med	IATA	No	0
110	Quam, M. Sessions, O. Kamaraj, U. *et al*	2016	Am J Trop Med Hyg; **94**; 2; 409-12	Japan National Tourism Organization	No	3
111	Quam, M. and A. Wilder-Smith,	2016	J Travel Med; **23**; 6	IATA	No	2
112	Read, J. Diggle, P. Chirombo, J. *et al*	2015	Lancet; **385**; 9962; 23-24	OAG (Traffic Analyser)	No	2
113	Rocklov, J. Quam, M. Sudre, B. *et al*	2016	EBioMedicine ; **9**; 250-6	IATA	No	2
114	Ruan, S. Wang, W. and S. Levin,	2006	Mathematical Biosciences and Engineering; **3**; 1; 205-218	IATA	No	1
115	Rvachev, L and I. LonginiJr,	1985	Math Biosci; **75**; 3-22	OAG; ICAO; Air Transportation Statistics; Australian International Arrivals; unknown	No	(1+1+0+1+0) 0.6
116	Sato, A. Sawai, H. Ito, I. *et al*	2015	Proceedings 2015 Ieee International Conference on Big Data	OAG	No	2
117	Schneider C, T Mihaljev, S Havlin, *et al*	2011	Phys Rev; **84**:6	Unknown	No	0
118	Semenza J, B Sudre, J Miniota, *et al*	2014	PLoS Neglected Trop Dis; **8**:12	IATA	No	0
119	Sessions, O. Khan, K. Hou, Y. *et al*	2013	Glob Heal Action ; **6**; 1	IATA ; OAG	Yes	(2+2) 2
120	Seyler, T. Grandesso, F. Le Strat, Y. *et al*	2009	Epidemics ; **1**; 3; 175-84	EuroStat; IATA ; ICAO	Yes	(1+0+0) 0.33
121	Struchiner C, J Rocklöv, A Wilder-Smith, *et al*	2015	PLoS ONE; **10**:8	Singapore Tourism Board	No	1
122	Tatem, A. Hay, S. and D. Rogers,	2006	PNAS; **103**; 16; 6242-6247	OAG	No	1
123	Tatem, A. and S. Hay	2007	Proc R Soc Lond B. Biol Sci; **274**; 1498-6	OAG (OAG MAX)	No	2
124	Tatem, A. Huang, Z. Das, A. *et al*	2012	Parasitology; **139**; 14; 1816-30	US Office of Travel and Tourism OAG	No	(2+1) 1.5

(Table 2.5 continues on next page)

(Table 2.5 continued)

Number	Authors	Year	Journal reference	Sources used	Validation	Reproducibility score *
125	Tatem, A. Rogers, D. and S. Hay,	2006	Malar J; 5; 57	OAG	No	1
126	Tian, H. Sun, Z. Faria, N. *et al*	2017	PLoS Negl Trop Dis; **11**; 8	ICAO	No	2
127	Tizzoni, M. Bajardi, P. Poletto, C. *et al*	2012	BMC Med; **10**; 165	IATA; OAG	Yes	(0+1) 0.5
128	Tuncer N and T Le	2014	Int J Crit Infr Prot; **7**; 27-47	US Department of Transport	No	2
129	Urabe, C. Tanaka, G. Aihara, K. *et al*	2016	PLoS ONE; **11**; 12	ICAO	No	1
130	Weinberger D, T Krause, K Molbak, *et al*	2012	Am J Epidemiol; **176**; 649-655	Icelandic Tourism Board; Statistics Iceland; Keflavik Airport	No	(4+3+2) 3
131	Wilder-Smith, A. and W. Leong,	2017	J Travel Med; **24**; 4	UNWTO	No	2
132	Wilder-Smith, A. Leong, W. Y. Lopez, L. *et al*	2015	BMC Med; **13**; 1	IATA	No	1
133	Wilder-Smith A, M Quam, O Session *et al*	2014	Euro Surveill; **19**:8	IATA	No	2
134	Wilson, S. Khan, K. Gilca, V. *et al*	2015	BMC Infect Dis; **15**; 1	IATA (Airport Intelligence Services – Passenger data)	No	1
135	Xiao L, H Zhang, Y Tang, *et al*	2015	SCRAMcon	OAG	No	1
136	Yoneyama, T. and M. Krishnamoorthy,	2012	SIMULATION; **88**; 4; 437-449	UNWTO; UNWTO	No	(1+1) 1

*Average total score shown with individual source scores shown in brackets.

Chapter 3 – Airline data description

Preamble

Understanding the data to be used in this thesis was crucial before any meaningful interpretation or analyses to be conducted could be undertaken. To the best of the author's knowledge, a detailed description of a closed source airline data set had not previously been attempted and/or reported. As shown in the previous chapter, a number of airline data sets from a variety of sources are available, but what each one represents may not always be clear. This chapter aims to provide a deeper understanding of the data itself.

Abstract

There is an understandably increasing pressure to ensure the data used in scientific research is accurate and fit for purpose, with publications such as Nature now requiring data set descriptors to be submitted with the accompanying manuscript. Fully understanding the data set that will be used for analysis and ensuring the absence of bias (or mitigating it) is crucial. As was shown in the previous chapter, a wide range of airline data set are available, each with its own advantages and drawbacks. The aim of this chapter was to describe in detail the OAG Traffic Analyser data set (sold as detailed airline bookings covering routings between global airports (including stopovers)) and uncover any trends and biases that may be present.

A series of methods were used to describe the data set, including aggregating airport bookings by country and regions, as well as temporal aggregations to quarters and years, and broken down to daily bookings. Airport bookings were also aggregated to country level and regional data and the number of bookings compared.

The downloaded data spanned from February 2010 to May 2015 at a monthly resolution and included 6,726 international airport codes and 12.77 billion bookings. Clear seasonal trends could be seen with peaks in July and August of every year with the overall number of bookings increasing each year. Some airports were used for varying travel purposes, such as Hartsfield-Jackson Atlanta International (ATL), an important connecting airport and Beijing Capital International (PEK), an important departing airport. As well as airports, the data also included 669 railway stations, bus and ferry terminals across 31 countries. Finally, a sudden increase in connectivity was also seen from 2014 onward.

Although sold as very detailed and accurate airline data, the collection methods for the OAG Traffic Analyser data remained undisclosed and therefore unclear. The data provides an understanding of the global airline network and its seasonality, however, it did not allow any understanding of the passenger demographics (age, sex, purpose of travel among others). Without having undertaken an in-depth analysis of the data set, the presence of railway stations and change in connectivity would not have been identified. Therefore, it is important to undertake a preliminary in-depth analysis of one's data before its use, followed by a validation with an independent, yet comparable data set, as done in the following chapter.

Introduction

Big data is being increasingly used in research, but no clear reporting guidelines allow for proper acknowledgments, hindering replication by other authors (Mooney and Newton, 2012). However, it is important that data used in any research is accurate and representative (Emanuelson and Egenvall, 2014) and that the readers are aware of what the data represents and how it was manipulated prior to use. Peer review journals such as Nature require authors to provide a data set descriptor that includes a number of mandatory fields such as data set name, authors and affiliations, abstract, background, summary and methods (data manipulation and access code), data records and validation (Nature, 2014). Additionally, having a standardised method of describing data would allow for a more uniform manner of interpreting results (Yale Law School Roundtable on Data and Code Sharing, 2010).

As all data sets contain errors; to reduce their impact on the results and arrive to the correct conclusions, it is important for researchers to find and understand these errors. Additionally, validating secondary data sets to ensure they are 'fit for purpose' and correct is crucial (Emanuelson and Egenvall, 2014). As the systematic review in **Chapter 2** demonstrates, there is a need to improve the description of the nature, sourcing and manipulation of information used for international human infectious diseases transmission modelling, as well as a need for data validation. In this spirit, a full description of the data set used throughout this thesis (OAG Traffic Analyser) will be given here, with a validation in the following chapter.

A wide range of airline travel data sources are available to use to model the international spread of infectious diseases, however, each data set has its drawbacks. Firstly, the data is likely to have been collected for commercial or governmental reasons rather than for scientific research. This may lead the data provider to ask for payment to access the data (commercial data) or the data may be geographically restricted to a particular government's borders. Additionally, commercial data providers may not be able to share their data collection methodologies for commercial reasons. It is therefore desirable to have a comprehensive understanding of what the data used represents and determine whether it is 'fit-for-purpose'. This may be considered an initial step in a transparent and reproducible methodology. It is also important to assess any possible bias that may be present in the data and understand the extent to which data manipulation by researchers prior to use may have contributed to biases.

The principal source of information described in this chapter is the OAG Traffic Analyser data, hereafter referred to as OAG. To the best of the author's knowledge, an extended description of this data set has not previously been undertaken and/or reported. Therefore, the aim of this chapter was to introduce the reader to the data set that will be used throughout this thesis. It is hoped that such data description will enlighten the reader regarding the use of certain variables used from OAG further on. A brief introduction about the company providing the data will start this chapter, before the data itself is described in detail. Finally, trends and biases discovered will be discussed.

About the data set and its provider

OAG was first implemented under the name of 'Official Airline Guide' with the first "Official Aviation Guide to The Airways" published in 1929 (OAG, 2016b). The company has since grown to be the "largest global airline network data provider", with over 25 million flight status updates made daily (OAG, 2016b). OAG aims to "connect the world of travel" by aggregating data from various sources and providing real time insights into the network. In keeping with this goal, the company claims to be the first provider of flight schedules and status as well as network analysis of flight data (OAG, 2016a). This information is sold to and used by airlines, airports and governmental agencies, as well as companies specializing in the travel industry (OAG, 2016a).

In 2013, OAG launched its Traffic Analyser database containing data on passenger traffic routes, to gather better data on numbers of passengers travelling, and predict future trends (OAG, 2013). The data available for download during the active license period (beginning August 2014) for this project ranged from February 2010 to one year in the future (i.e. July 2016) (OAG, 2016c). Finally, the number of adjusted bookings ("Bookings.Adjusted.") represents the "true total market figures" according to (2015). These numbers result from a "sophisticated algorithm" using data from OAG's own schedule database and the 'passenger traffic' from the Global Distribution System (GDS) and other sources. The unadjusted number of bookings ("Bookings.Unadjusted.") are the average fare from Travelport tickets (OAG, 2017), however, the exact differences between the data sets were not shared with us when enquired.

It was clear when directly comparing the number of adjusted and unadjusted bookings by year and month (**Figure 3.1**) that these were very different numbers. Firstly, the Unadjusted bookings were absent from the data until January 2011 and there was an absence of seasonal trends. The author had an a priori knowledge of the seasonal trends present in airline travel, which was present in the Adjusted bookings making the case for using the Adjusted bookings rather than the Unadjusted.

Reasoning for choice in data provider

The use of these data was recommended by other modelling groups who had previously used airline travel data to model the 2009 'swine flu' AH1N1pdm09 influenza pandemic spread (Fraser *et al.*, 2009). Enquiries with OAG in 2014 suggested the data they could provide would be suitable for modelling purposes relating to exportation of Ebola virus disease from West Africa during the 2014-2015 epidemic. The funders of this thesis, Health Protection Research Unit in Emerging and Zoonotic Infections (HPRU EZI) were specifically interested in modelling the risk to other countries posed by the outbreak.

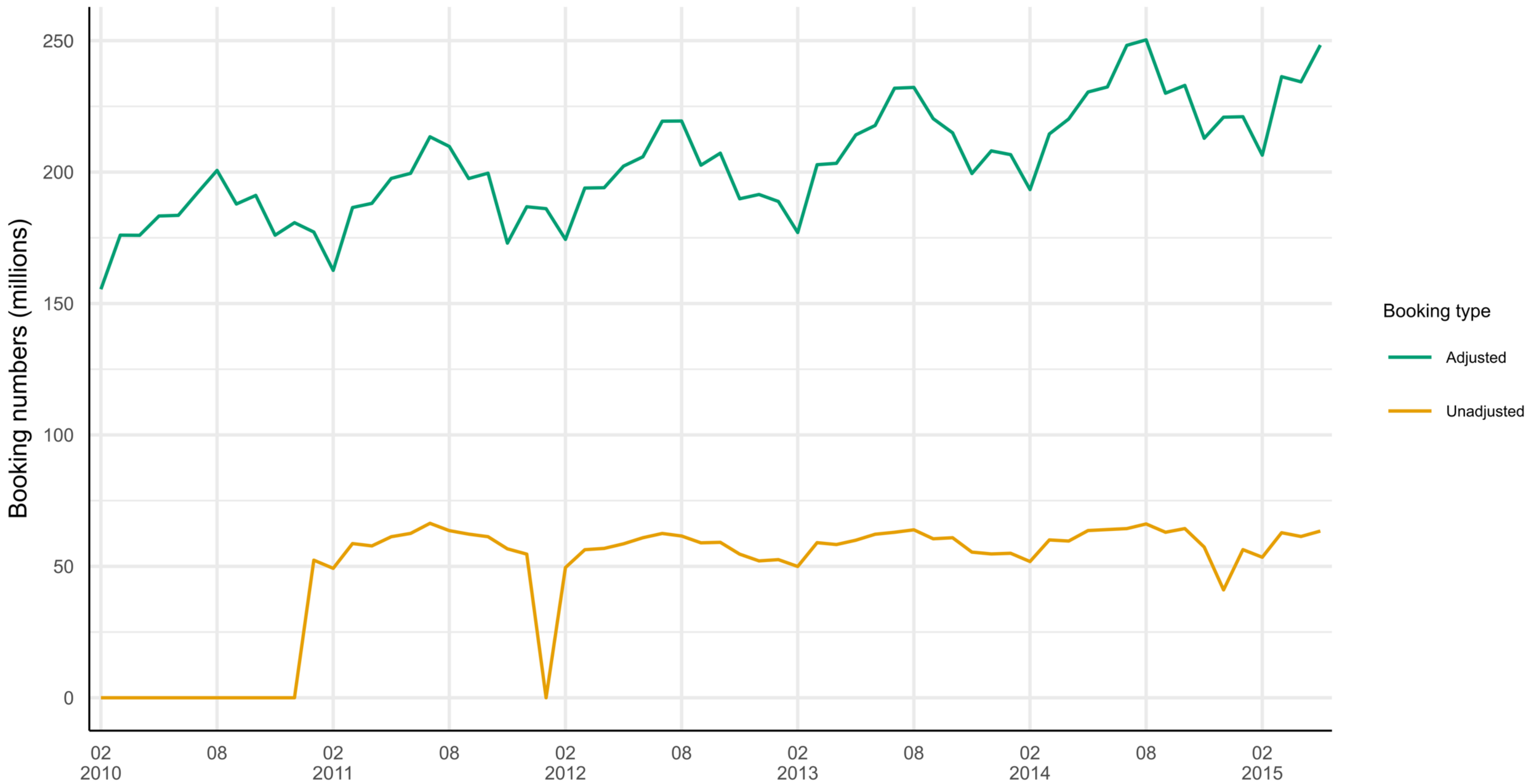


Figure 3.1: Direct comparison of OAG's Bookings.Adjusted against Bookings.Unadjusted. by month and year.

Data description

The OAG Traffic Analyser data set (thereafter referred to as OAG) was downloaded over the course of one year (August 2014 to July 2015) from www.oag.com, following the data provider's licensing restrictions. The downloaded data ranges from February 2010 to May 2015 (a total of 64 months), as January 2010 was not available in the database to download and it was thought that June 2015 was too close to the collection date to be truly representative, as the company updates their data monthly so was not downloaded. The final monthly updates made by the company occur on the third weekend of every month and cover the latest four months (OAG, 2017). Therefore, when downloading in July, it is possible the months of March to June were still being updated, with June being the most likely to see changes in these updates. Downloading bookings for future months was thought not to be a true representation of bookings for the same reasons. The 'Original-Destination' database from Traffic Analyser offers data regarding flights on global, national and airport level. For the purposes of this thesis, when downloading the data an origin was always determined (continental, national or state level), but without specifying a destination. This ensured that OAG would return information on bookings departing from the desired point of origin, with a destination anywhere in the world (international and domestic flights alike). Although not detailed in the original data, we were able to determine the direction of travel by knowing the booking's departing airport code and the trip point of origin (see definitions below). Both variables were downloaded for each temporal and geographic resolution. **Figure3.2 A** shows the OAG regions available in the data set, as defined by the company. In this context, regions are defined as neighbouring countries. Given the data file sizes and the downloadable file size restrictions, the USA and Western Europe had to be broken down into smaller groups to be downloaded (**Figure 3.2B** and **C**). On the other hand, regions with few bookings (such as Africa) were grouped and downloaded together. These regional groupings were only used to download data from September 2014 onward (due to downloadable file size restrictions). Once all regions and countries were downloaded, they were collated to form a monthly data set of all global bookings.

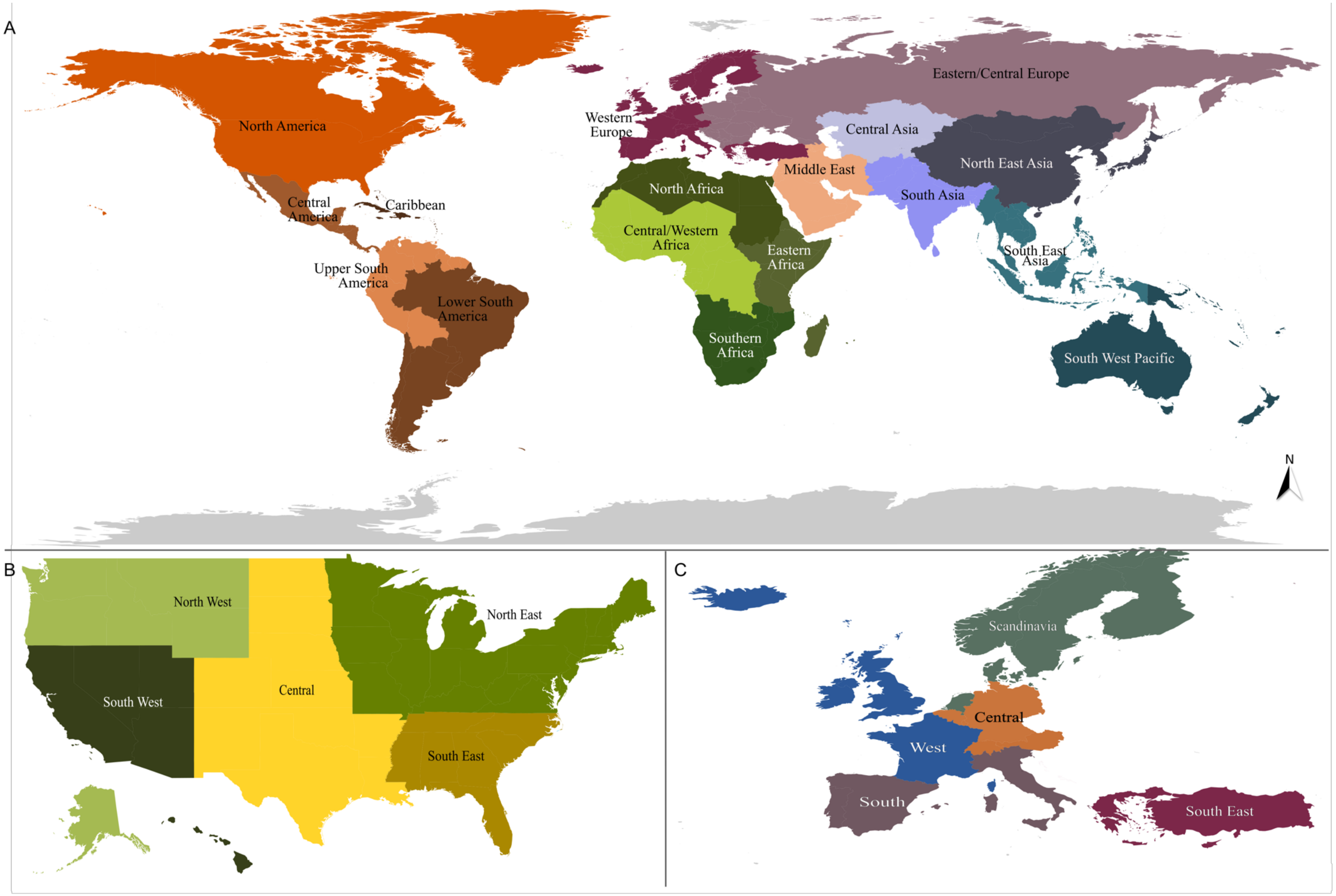


Figure 3.2: Representation of world regions as defined by OAG (A), groupings of states of the United States of America (user-defined) (B) and of Western European countries (user-defined) (C).

The Traffic Analyser data base offers the client the possibility of tailoring the reports according to their needs (OAG, 2013). The fields chosen for download were:

- **“TimeSeries”**: Annual or monthly resolution available, between 2010 and 2015. Data were retrieved at the finest temporal resolution available, monthly, to gain a detailed idea of seasonal patterns. This also allowed quarterly or annual aggregation of bookings, if needed.
- **“Routing”**: the complete and ordered routing (according to flight itinerary), including departing, arriving and connecting airports, presented as three character airport codes, known as International Air and Transport Association (IATA) codes. This included a maximum of two connections.
- **“Bookings.Adjusted.”**: These represent the number of bookings made per routing, but does not detail the number of passengers included in each booking. According to OAG (2016c) the “Bookings.Adjusted.” represent the true market figure (as defined by OAG), and were therefore used in our analysis. Unless stated otherwise, these will be referred to as “bookings” throughout the rest of this thesis.
- **“Bookings.Unadjusted.”**: Although a less accurate count of bookings (OAG, 2015), and resulting from models, “Bookings.Unadjusted.” were also recorded when downloading the data. This was not used in any of the analyses conducted in this thesis using the OAG data.
- **“Point.of.Origin.Cd”**: The origin airport for the bookings, not necessarily the origin for the routing for which the booking is returned. This may be in a different city, region or country to the departing (origin) airport. For example, a routing of LHR-JFK may have LHR or JFK as a point of origin depending on whether the booking originated in the United Kingdom or the USA, respectively, or the point of origin may be a third airport if this is part of multi-stop journey. This information helped determine the journey type (inbound or outbound) and whether it was domestic or international, which in turn also helped determine any directional trends.

A brief description of the data format can be found in the Data Dictionary at the end of this chapter. Each file was downloaded in CSV format and manipulated using the open access software R, version 3.4.1 (www.R-project.org). Once all the monthly data sets were downloaded and corresponding monthly data sets collated together, data cleaning and manipulation was done as follows to generate distinct sets of data:

- **Aggregation into quarterly data**: monthly data sets for each year were grouped as follows: January-March (Q1), April-June (Q2), July-September (Q3) and October-December (Q4). Unfortunately, Q1 of 2010 and Q2 of 2015 could not be aggregated as January 2010 and June 2015 could not be downloaded from OAG. A total of twenty quarters were thus generated, ranging from Q2 2010 to Q1 2015.
- **United Kingdom international routing data**: the number of international bookings arriving into and departing from the United Kingdom was generated per month and quarter. The data sets were aggregated by selecting routings with a point of origin in the UK and an arriving or departing airport in the UK, depending on the journey leg wanted (inbound international, outbound international).
- **Country level, origin-destination matrices**: origin-destination matrices showing the number of aggregated bookings were generated linking all countries present in the data sets. UK centric matrices were also generated using arriving and departing countries only (i.e. not considering the routing's point of origin).

The data were made up of a total of 6,726 airports and heliports each with its own IATA code, across 233 unique countries and territories. The United States was recorded as having the largest number of airport codes (n=1,135), followed by Canada and Australia, with 456 and 281 airports respectively, whereas the United Kingdom was recorded as having 168 airport codes. **Figure 3.3** represents the location of all airports present in the OAG data with the 20 busiest (in terms of aggregated passengers arriving, connecting and departing bookings) highlighted by their IATA codes, and listed in **Table 3.1** (ranked by size of passenger flow).

The total international connections between the ten busiest countries in terms of aggregated departing, connecting and arriving bookings is shown in **Figure 3.4**. Each colour corresponds to a departing country, and the width of each link representing the number of bookings associated to each routing (direct and indirect). Only international destinations were represented here as the number of domestic bookings was overwhelmingly larger, and is of less interest in the context of a pandemic. Looking at the United Kingdom, a total of 483.3 million bookings were recorded arriving and departing, between February 2010 and May 2015 the highest number of international passengers recorded. Spain was the most popular destination for UK passengers, with 88.1 million bookings, followed by the United States (39.0 million bookings). The UK was the first destination for passengers arriving from Spain (87.6 million bookings). The United States (country with the second highest number of international departures and arrivals recorded at 426.1 million bookings) had strong links with its neighbour Canada (73.3 million departing bookings), followed by the UK (40.1 million departing bookings). When aggregating the total bookings between the ten busiest countries by season, shown in **Figure 3.5**, some variations can be seen with Australia seeing the most departures in autumn and winter (corresponding to their summer), whereas the UK saw the most in spring and summer. The UK saw most international departures to Spain in the summer and the least in winter.

Representing the country to country connections in a heat map was difficult and could not be presented in a legible manner. However, **Figure 3.6** shows the regional level of connections, both within and between regions (defined here as groups of neighbouring countries). It can be noted that three regions dominate the internal number of bookings globally: North America (3.3 billion bookings), Western Europe (2.9 billion bookings) and North East Asia (2.6 billion bookings). The highest recorded number of inter-regional bookings was from Western to Eastern Europe, with 240.2 million bookings.

Table 3.1: List of the twenty busiest (largest aggregated departing, connecting and arriving bookings) airport codes globally with corresponding name, associated city and country as well as total number of bookings and corresponding percentage of global bookings ranked by size (%) of global bookings.

IATA code	Airport name	City	Country	Bookings (millions)	Global bookings (%)	Rank
PEK	Beijing Capital International	Beijing	China	364.32	1.27	1
LHR	Heathrow	London	United Kingdom	309.06	1.08	2
ATL	Hartsfield-Jackson Atlanta International	Atlanta	USA	308.03	1.07	3
LAX	Los Angeles International	Los Angeles	USA	283.40	0.99	4
HND	Haneda	Tokyo	Japan	279.92	0.98	5
ORD	O’Hare International	Chicago	USA	257.53	0.90	6
CGK	Soekarno-Hatta International	Jakarta	Indonesia	252.11	0.88	7
HKG	Hong Kong International	Hong Kong	Hong Kong SAR China	250.22	0.87	8
CDG	Charles de Gaulles	Paris	France	240.43	0.84	9
JFK	John F Kennedy International	New York	USA	233.06	0.81	10
DXB	Dubai International	Dubai	United Arab Emirates	231.93	0.81	11
SIN	Changi	Singapore	Singapore	227.25	0.79	12
CAN	Baiyun International	Guangzhou	China	220.33	0.77	13
BKK	Suvarnabhumi	Bangkok	Thailand	216.54	0.75	14
PVG	Pudong International	Shangai	China	215.86	0.75	15
DFW	Dallas/Fort Worth	Dallas	USA	215.23	0.75	16
FRA	Frankfurt International	Frankfurt	Germany	208.94	0.73	17
DEN	Denver International	Denver	USA	201.19	0.70	18
LAS	McCarran International	Las Vegas	USA	194.07	0.68	19
MAD	Adolfo Suarez-Barajas	Madrid	Spain	193.25	0.67	20

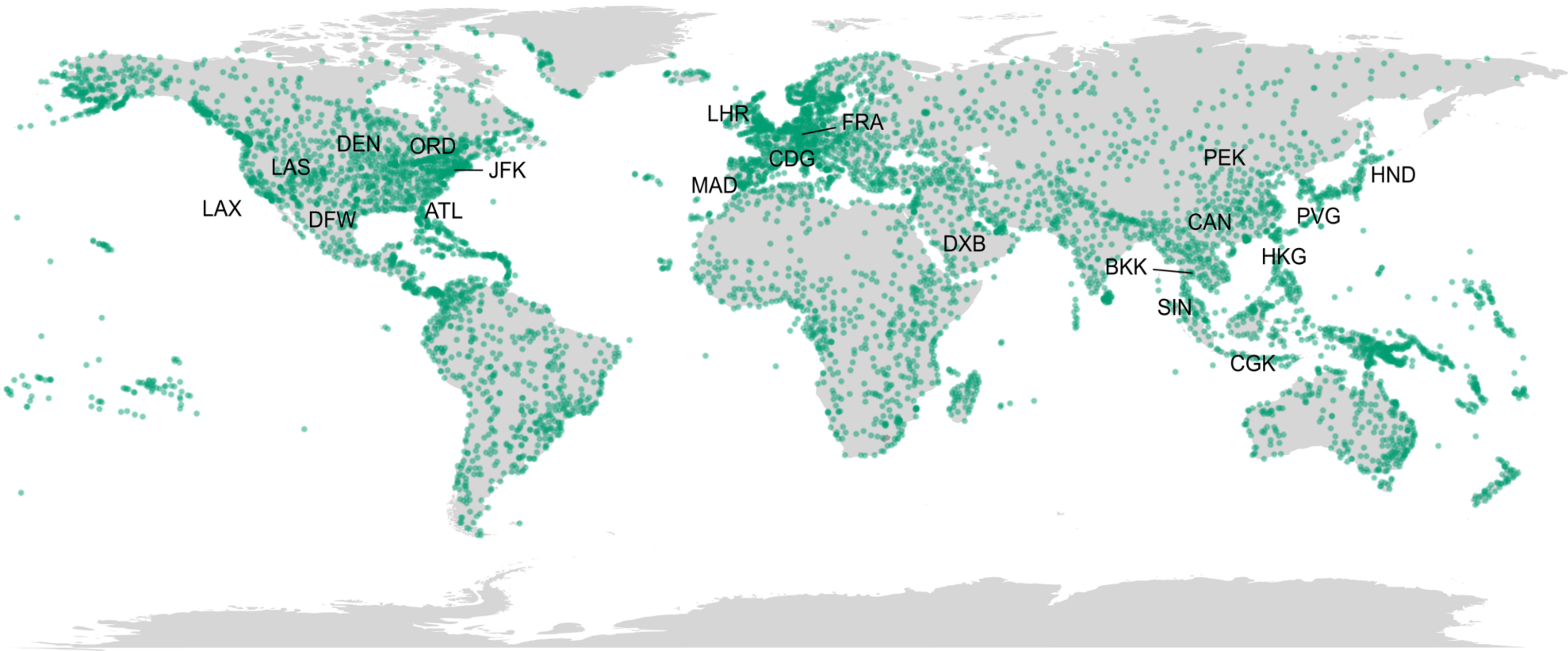


Figure 3.3: World map of all airports in the OAG Traffic Analyser data set, with the twenty busiest airports (most combined departing, connecting and arriving bookings) highlighted by their IATA codes.

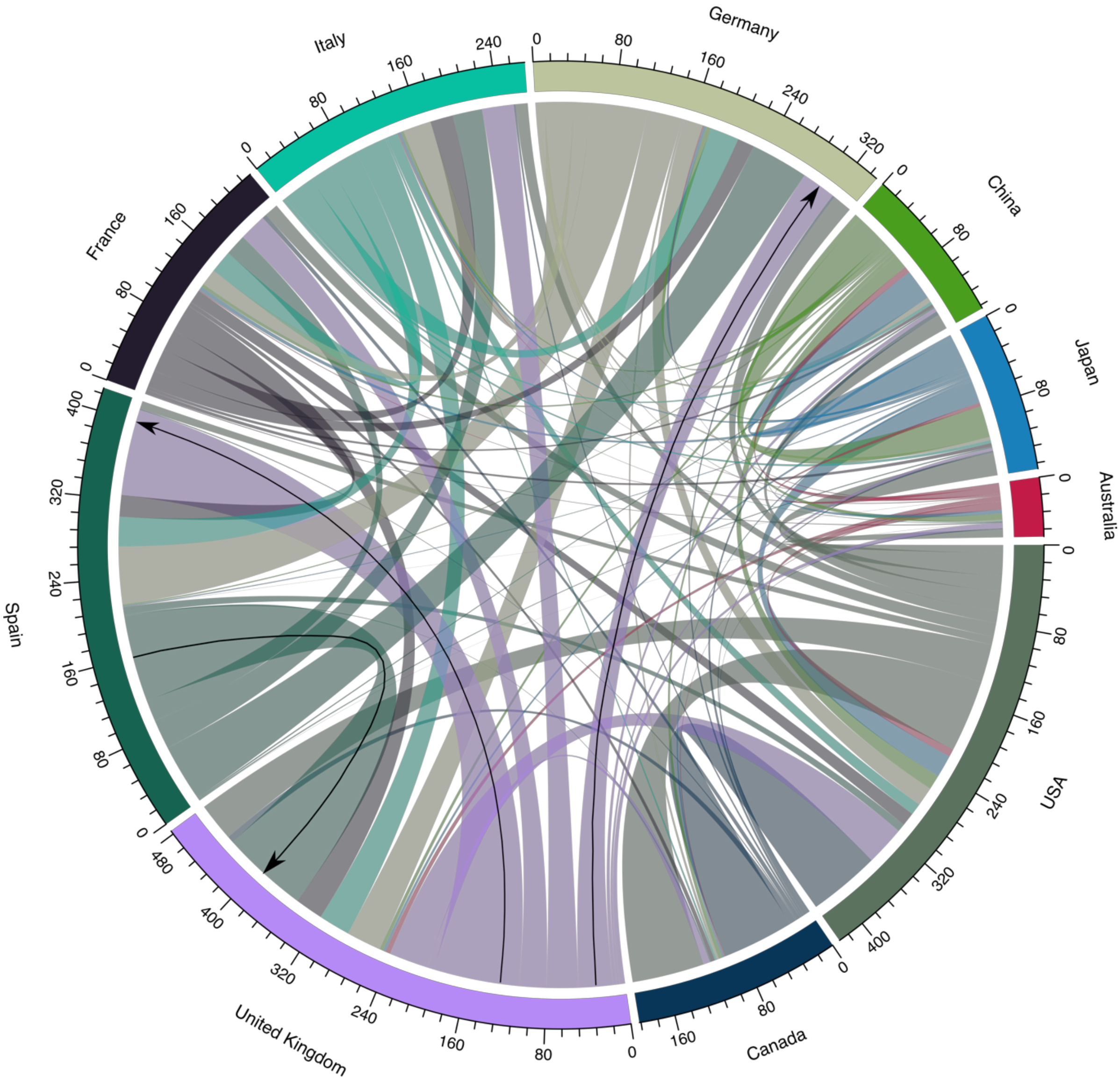


Figure 3.4: Connections (in millions) between the ten countries with the highest number of total cumulative international departing, connecting and arriving bookings, between February 2010 and May 2015.
Note: the arrows illustrate the direction of flow.

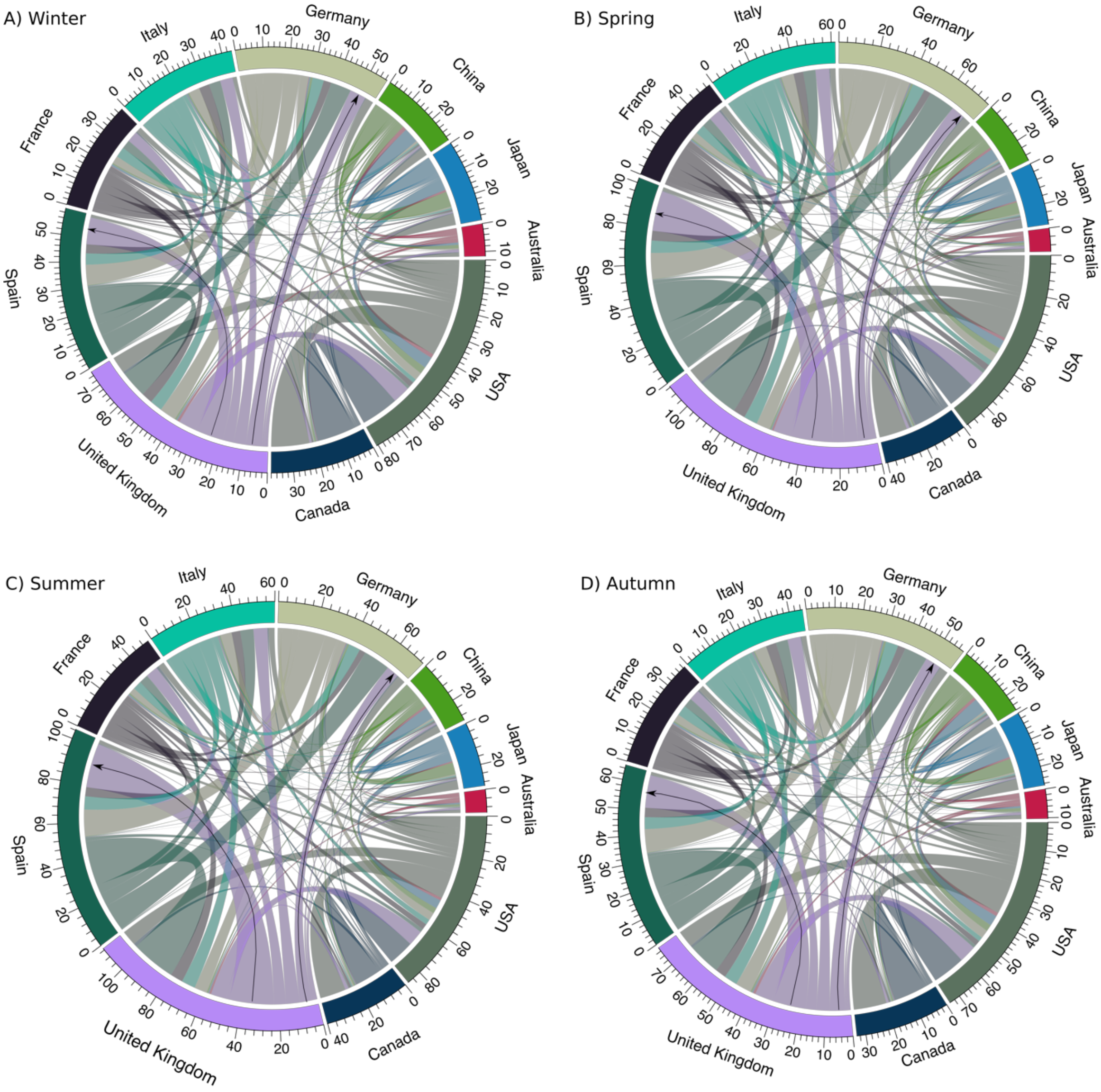


Figure 3.5: Connections (in millions) between the ten countries with the highest number of cumulative international departures, connections and arriving bookings, according to seasons: A) winter (January to March), B) spring (April to June), C) summer (July to September) and D) Autumn (October to December), between February 2010 and May 2015. Note: the black arrows illustrate the direction of booking flow.

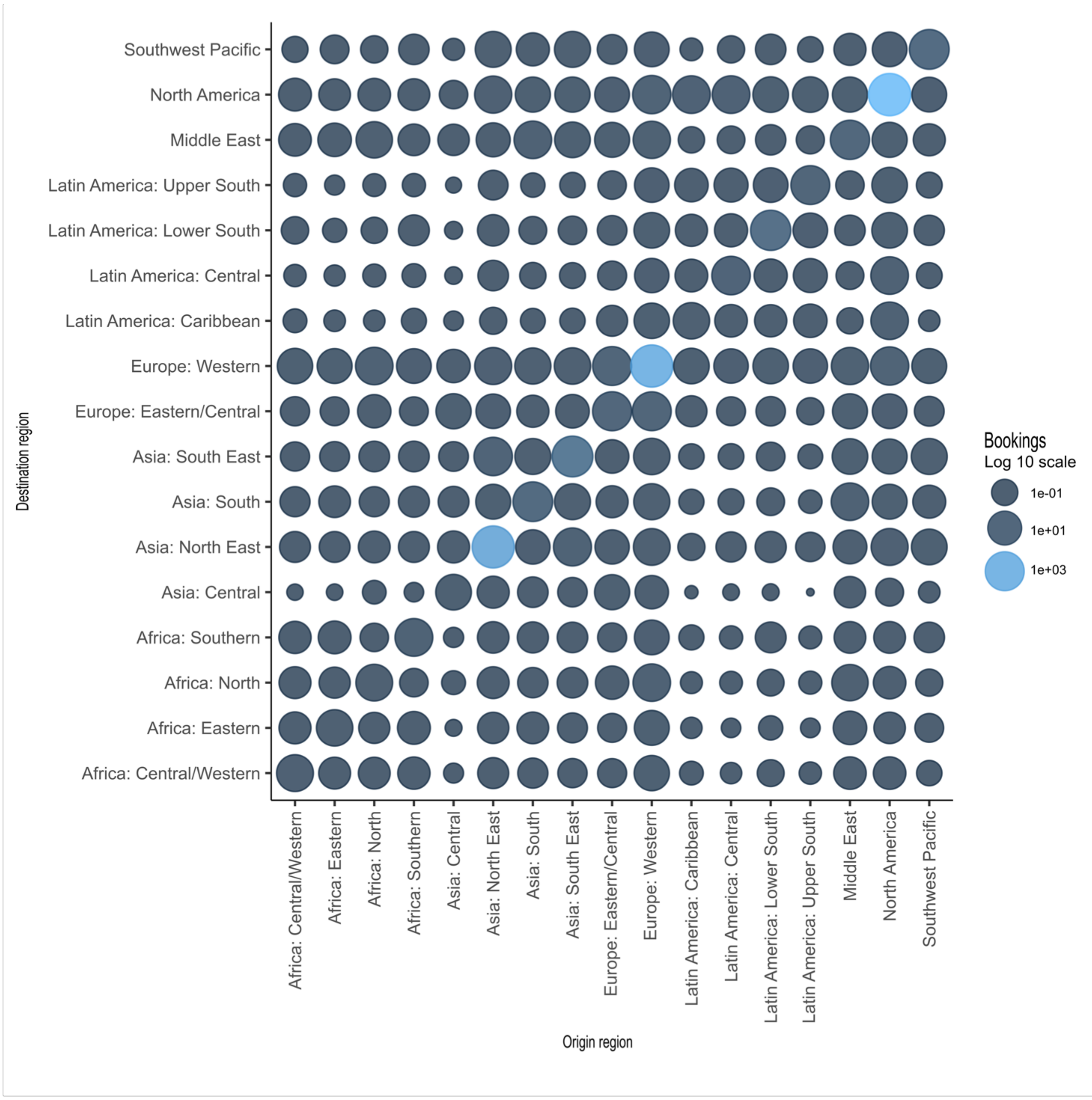


Figure 3.6: Heat map of bookings (millions) between and within OAG regions.

A total of 12.77 billion bookings were present in the data, between February 2010 and May 2015. An increasing trend was evident across the years, with distinct seasonal patterns: large peaks in July and August and troughs in February (**Figure 3.7 A**). An additional although smaller, peak could be seen in December of every year. The temporal seasonality was also considered according to the journey leg (**Figure 3.7 B**). The same seasonal patterns could be seen as in **Figure 3.7 A** but showed a large variation in number of connecting rather than departing or arriving bookings. The number of departing and arriving bookings overlay each other exactly as every flight that departs must arrive.

When aggregating the bookings by month (**Figure 3.8 A**), this increasing trend was still present with the number of bookings for each month increasing year on year, except for November, where there was a drop in number of bookings between 2010 and 2011 (175.4 million and 172.4 million bookings, respectively). August saw the largest number of bookings every year, with August 2014 recoding the largest number of any month with 250 million bookings, whereas the smallest number of bookings was seen in February 2010 with 155 million bookings.

When comparing the number of bookings per day to the bookings per month (F**igure 3.8 B**), the same seasonal patterns could be seen as in **Figure 3.8 A**. The fall in number of bookings was clearly seen in November 2011. The seasonality was not as clearly marked when looking at the average daily bookings by month and year (**Figure 3.8 B**) with the February dip previously seen, had now been flattened out. However, July and August still remained the two months with the largest number of daily bookings. An overall average of daily bookings was done with the four years where 12 months of data were available (2011 to 2014) (**Figure 3.8 C**), which also reflected the strong summer seasonality.

(Figure 3.7 continues on next page)

(Figure 3.7 continued)

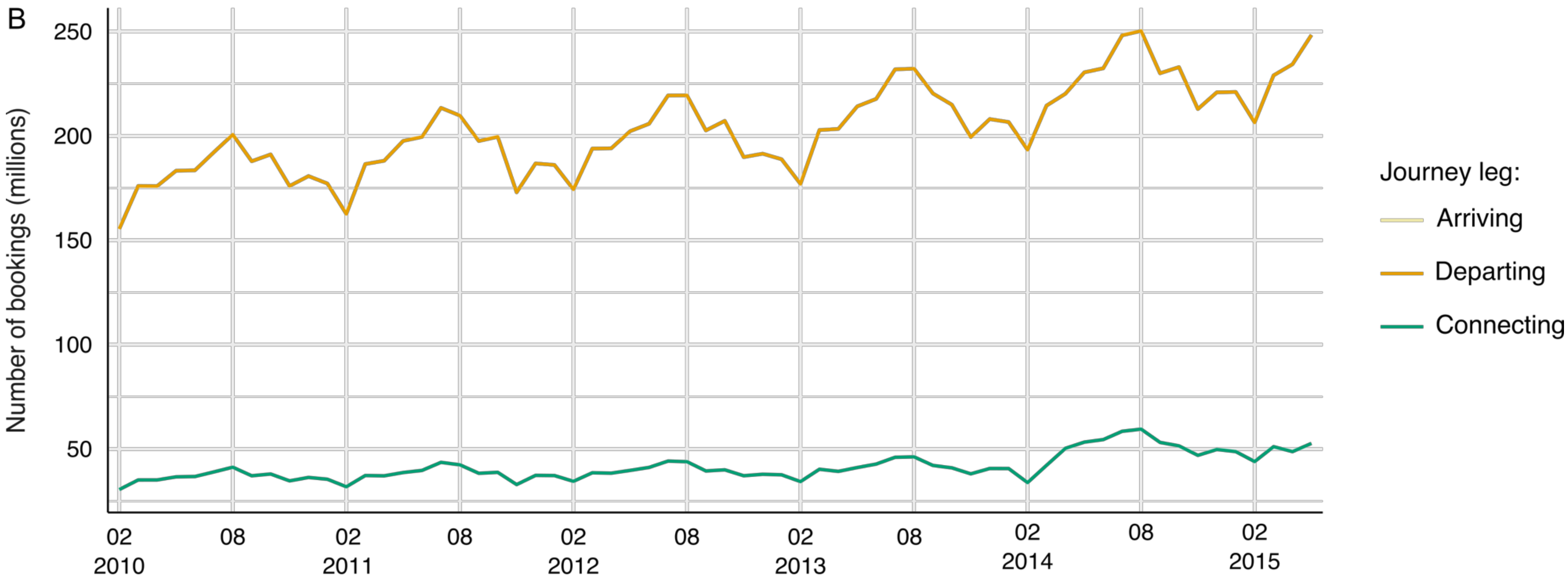
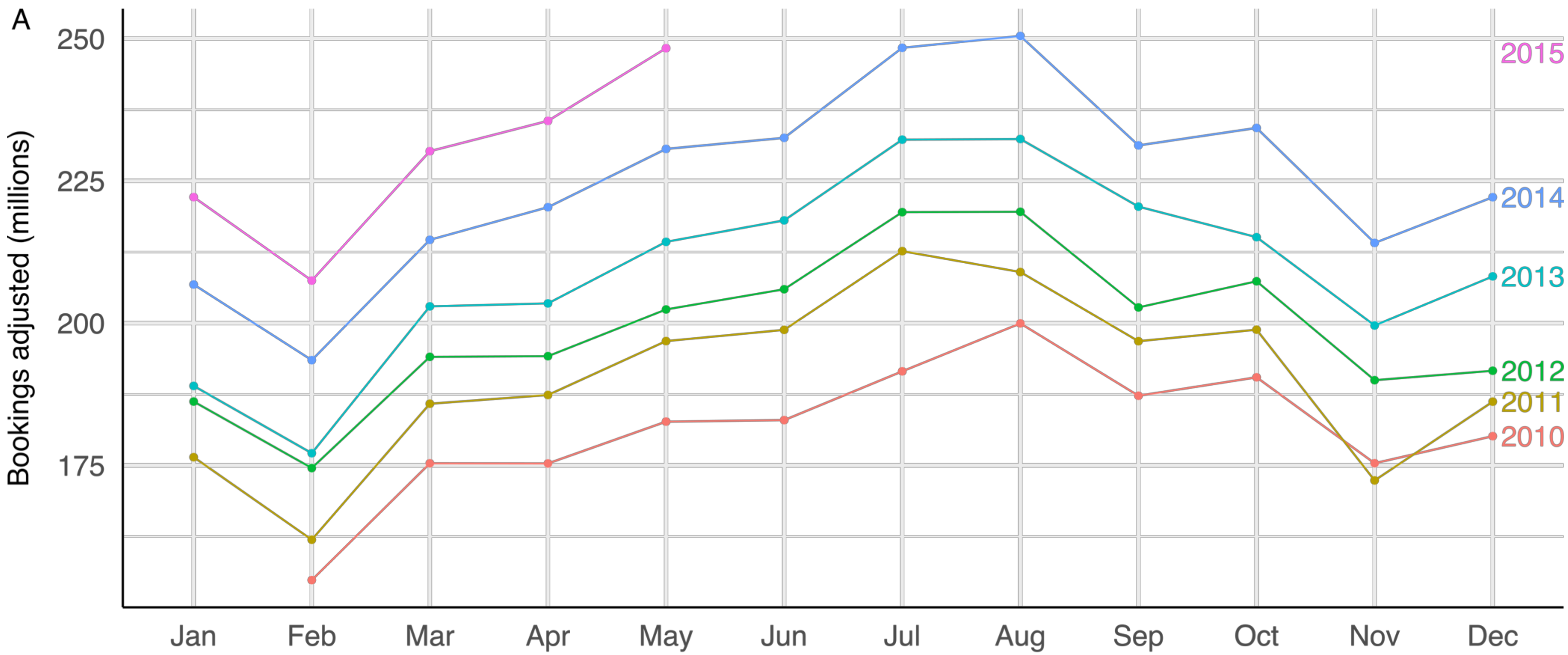
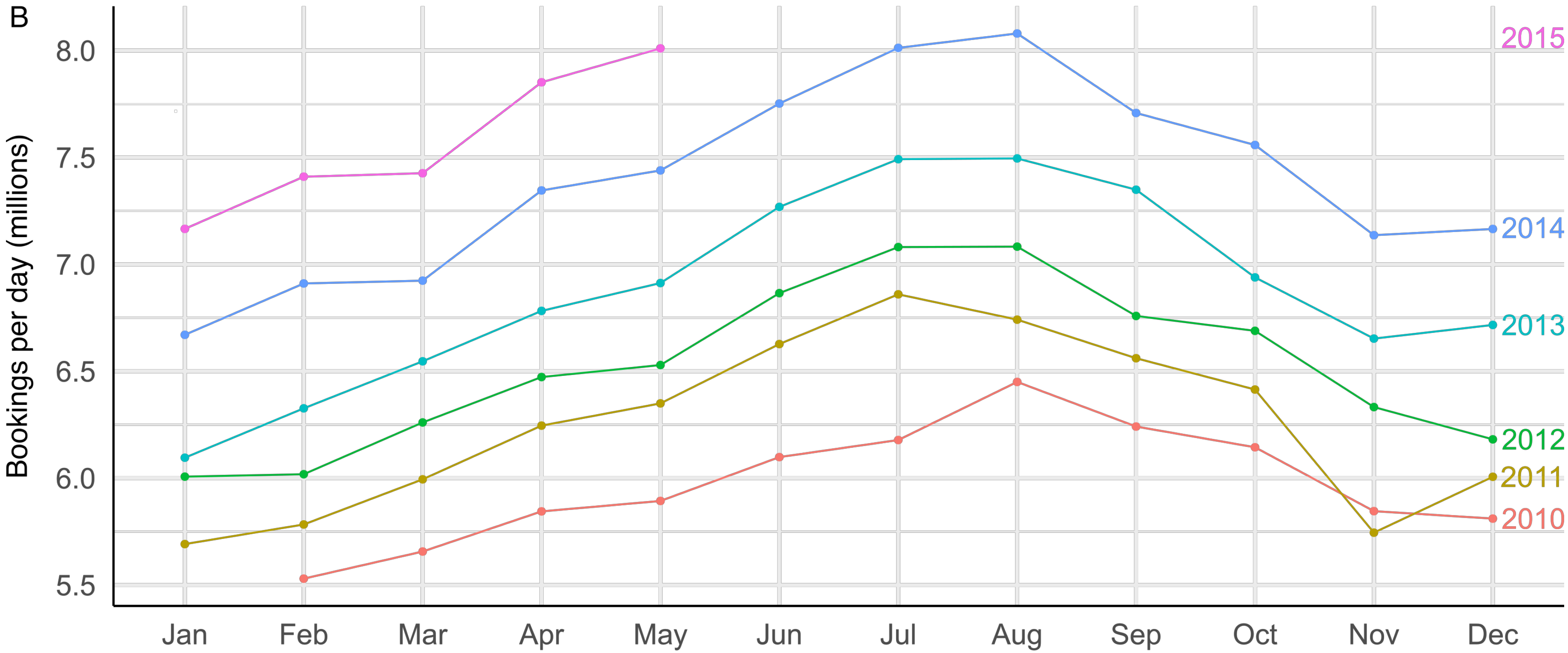


Figure 3.7: Total number of global bookings present in OAG by month (A); total number of bookings separated according to journey leg (departing, connecting and arriving) globally according to time (B), between February 2010 and March 2015.
Note: the sum of arriving and departing bookings overlay each other.



(Figure 3.8 continues on next page)

(Figure 3.8 continued)



(Figure 3.8 continues on next page)

(Figure 3.8 continued)

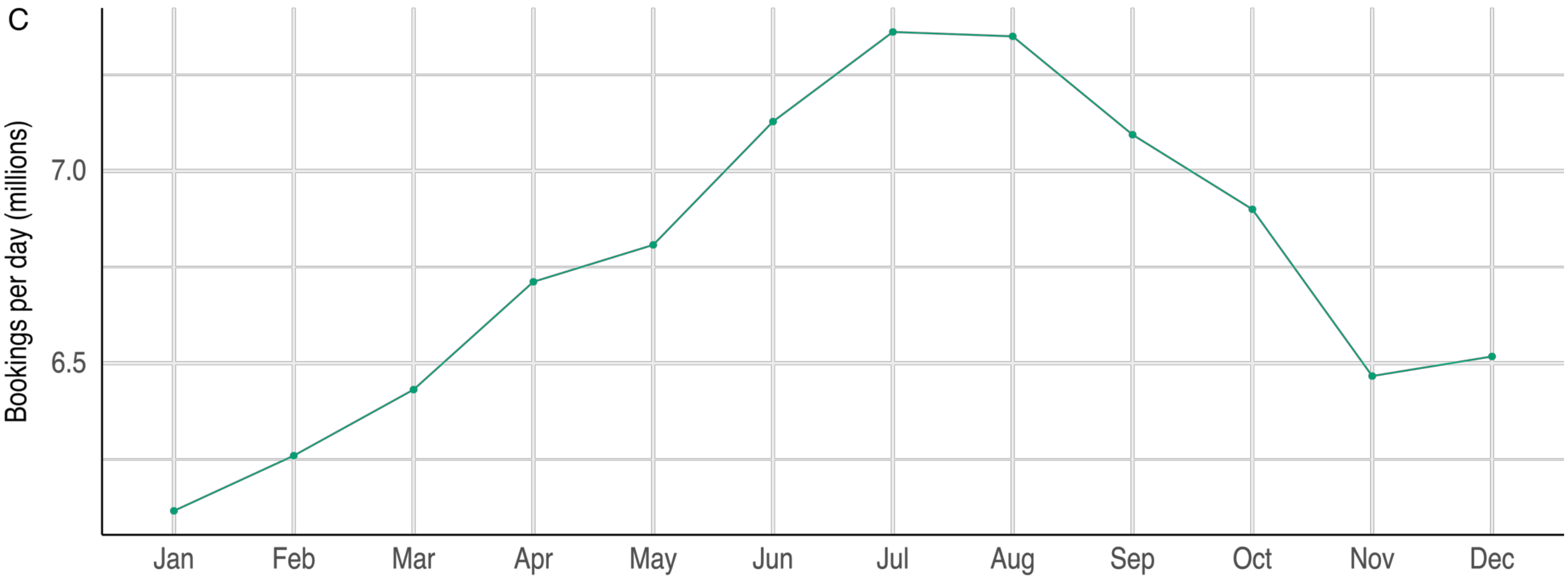


Figure 3.8: Monthly bookings aggregated by month and year (A); daily bookings aggregated by month and year (B) and average daily bookings by month, for years containing 12 months of data (2011 to 2014) (C).

Airports and countries acting as sinks and sources

To determine whether any bookings had been missed from any airport during the downloading stage, and to determine any initial trends in the data, a sinks and sources analysis was done. By separating each routing into the individual airport codes, and determining the number of departing (B_{dep}), connecting (B_{con}) and arriving bookings (B_{arr}), any airport with a bias in their number of B_{dep}, B_{con} or B_{arr} was easily identified. Manipulating and visualizing the data in the following way allowed for an initial understanding of the data structure, any gaps and/or trends present. This analysis also allowed for an initial understanding of the role of various airports in the network, such as which were of most importance (largest number of total bookings), or had most B_{con}. It was also possible to identify which airports had more arrivals (sinks) or departures (sources), and determine any seasonality present within the data.

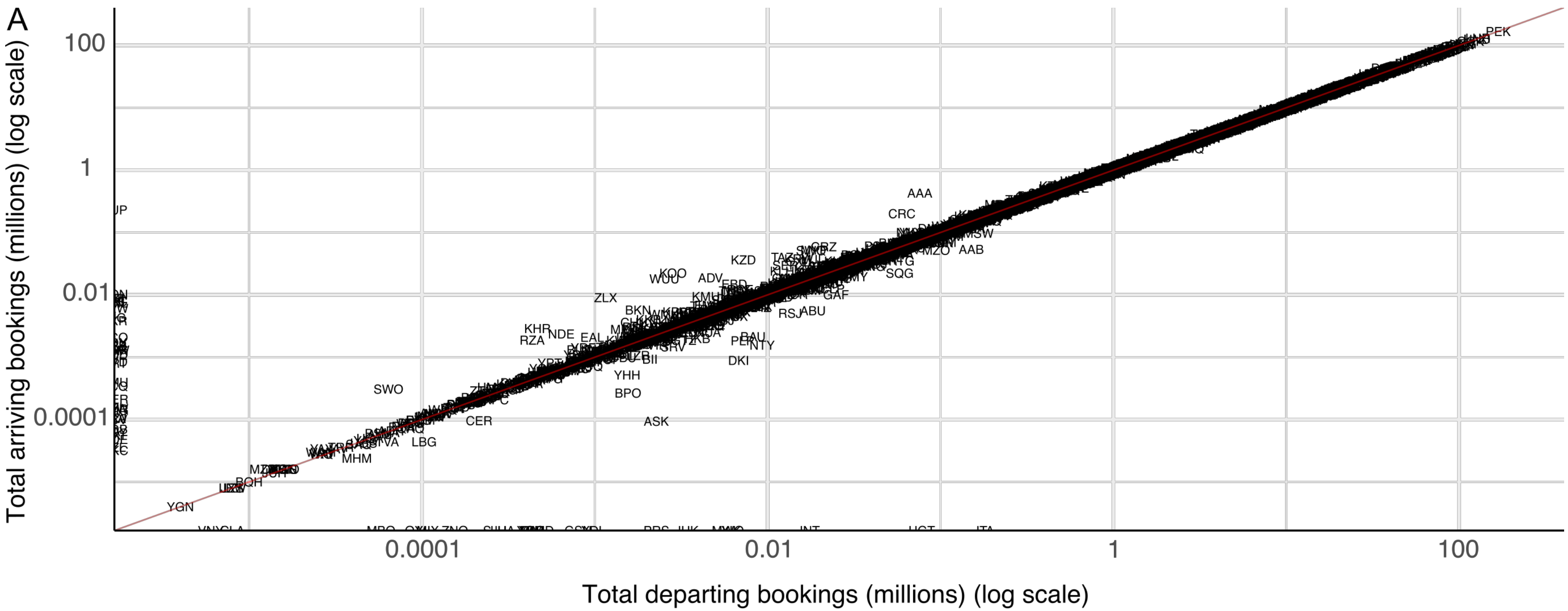
Ignoring the point of origin, each routing was broken down to individual component airport and assigned one of three categories: departing, connecting or arriving, according to its location within the routing thread. For example, a routing of LHR-JFK-LAX, had LHR and LAX as departing and arriving airports respectively, and JFK as a connecting airport. The number of bookings for each airport was then carried over and aggregated according to their category (arriving, departing or connecting). This was performed for each time resolution available, resulting in a data set with the cumulative number of bookings for each airport. A global summary was plotted against time to see any global patterns, shown in **Figure 3.7 B**. The total journey leg bookings were then summed by airport and plotted on a natural log scale (**Figure 3.9 A**). A number of airports could be seen to have large differences in the number of bookings recorded either departing or arriving. These airport codes included railway and bus stations through which no bookings were recorded. However, 37 airports had no departing bookings (but recorded up to 229,762 bookings accumulated across all time periods) and 22 airports had no arriving bookings but recorded 181,100 departing bookings. Beijing Capital International airport (PEK) was the busiest airport with the largest number of passenger bookings (combined arriving and departing bookings) (335.39 million bookings), followed by London Heathrow airport (LHR) (259.3 million bookings) and Haneda airport (HND) (252.0 million bookings) (**Figure 3.9 A**).

There was good agreement between the number of departing and arriving bookings for each airport, except for Washington Dulles International airport (IAD) and Ronald Reagan

Washington airport (DCA), where graph **Figure 3.9 B** shows more departing than arriving bookings for both airports. When comparing all bookings (including connections) to the net difference (B_{dep} - B_{arr}) of bookings, Hartsfield-Jackson Atlanta International Airport (ATL) became the third busiest airport globally (308.0 million bookings), most likely due to the large number of B_{con} passing through the airport (158.2 million connecting bookings). Given the high number of connecting bookings linking ATL to other airports (51% of bookings), this airport was a good example of a hub airport. Hubs are defined here as an airport with an important number of connections to other airports and that is central to the airline network (Woolley-Meza *et al.*, 2011). Airport centrality is defined as having a high number of geodesic paths passing through, connecting distant parts of the network (Newman, 2003). When determining which airports act as sinks ($B_{arr} > B_{dep}$) and sources ($B_{arr} < B_{dep}$) (**Figure 3.9 B**), it became clear that London Heathrow (LHR), Adolfo Suarez Madrid-Barajas (MAD) and Amsterdam Schiphol (AMS) airports act as important sinks, with $B_{arr} > B_{dep}$. On the other hand, Incheon International airport (ICN), Haneda (HND) and Beijing Capital airport (PEK) act as sources, with $B_{arr} < B_{dep}$.

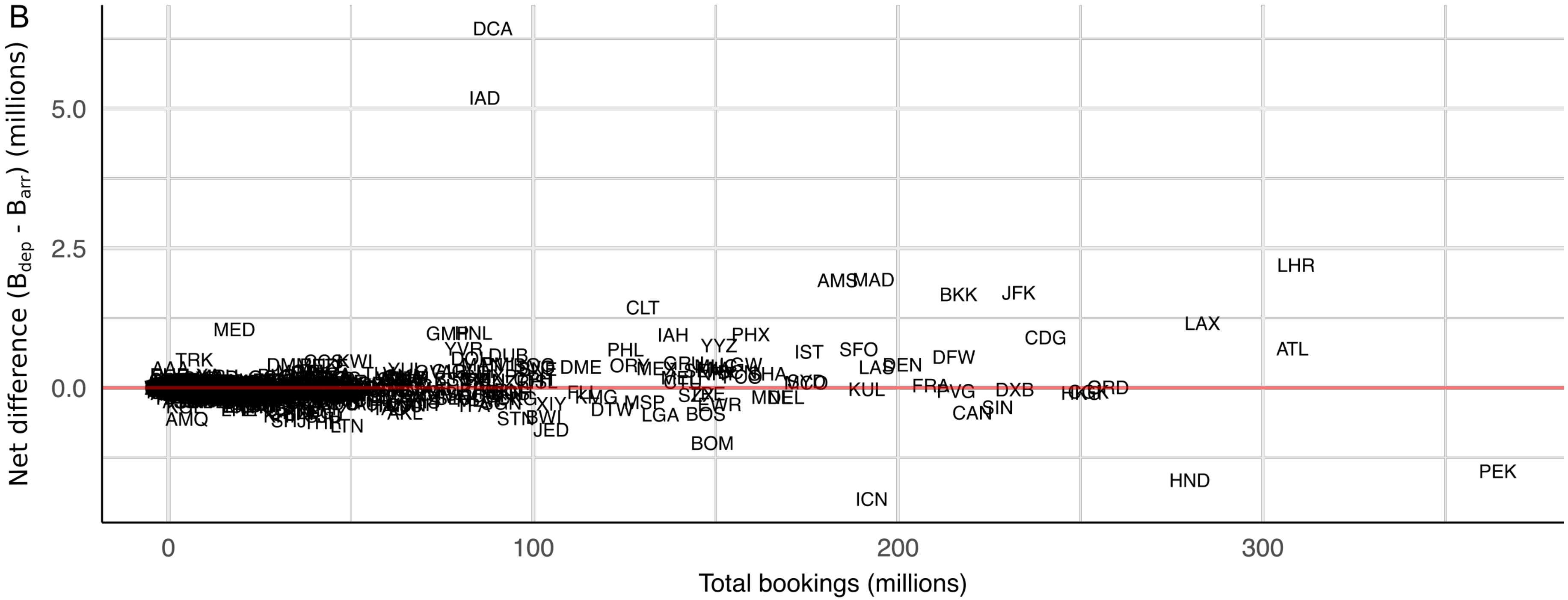
However, two airports stood out from this analysis: Washington Dulles International (IAD) and Ronald Reagan Washington (DCA) airports, because of missing departing data (departures from these airports were recorded as zero bookings, which is extremely unlikely), an error made during data download. These airports are located in Washington DC within the OAG database, which is considered as an independent US state (District of Columbia) in the data base, therefore were missed when downloading data at the state level. This error occurred for data relating to September 2014 onwards. This missing data were addressed by assuming that the number of departing bookings and routings match exactly the number of arriving bookings and the routings recorded and travel type were reversed from the data already available. The decision was taken after considering whether other important airports (such as JFK) showed similar patterns. It was also assumed that DCA and IAD had no connecting bookings during this time period.

When identifying connecting airports (**Figure 3.9 C**), plotting the total number of connecting bookings (B_{con}) against the total difference (($B_{arr} + B_{dep}$) – B_{con}) showed that only two airports are mostly used for connections: Charlotte Douglas International airport (CLT) (79.43 million connecting bookings) and Hartsfield-Jackson Atlanta International airport (ATL) (158.16 million connecting bookings). Interestingly, Beijing Capital International airport (PEK) has much fewer connections than arriving or departing bookings (28.93 million connecting bookings).



(Figure 3.9 continues on next page)

(Figure 3.9 continued)



(Figure 3.9 continues on next page)

(Figure 3.9 continued)

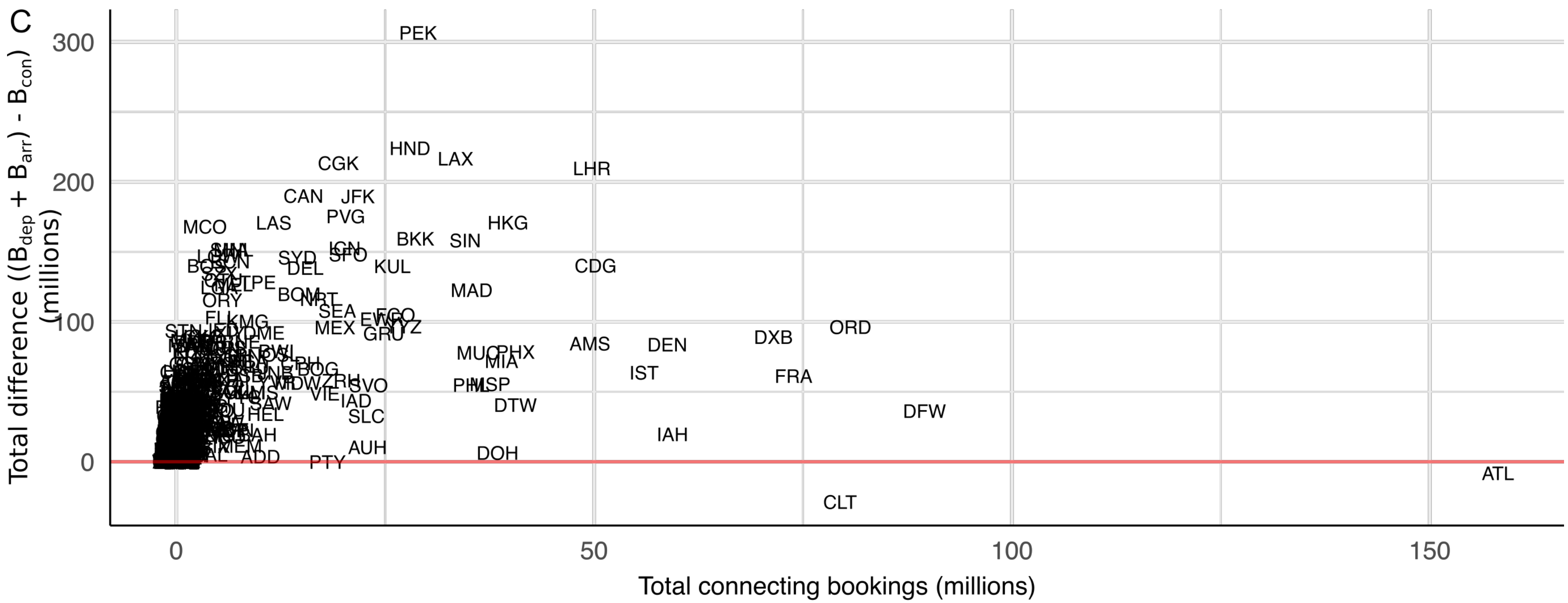
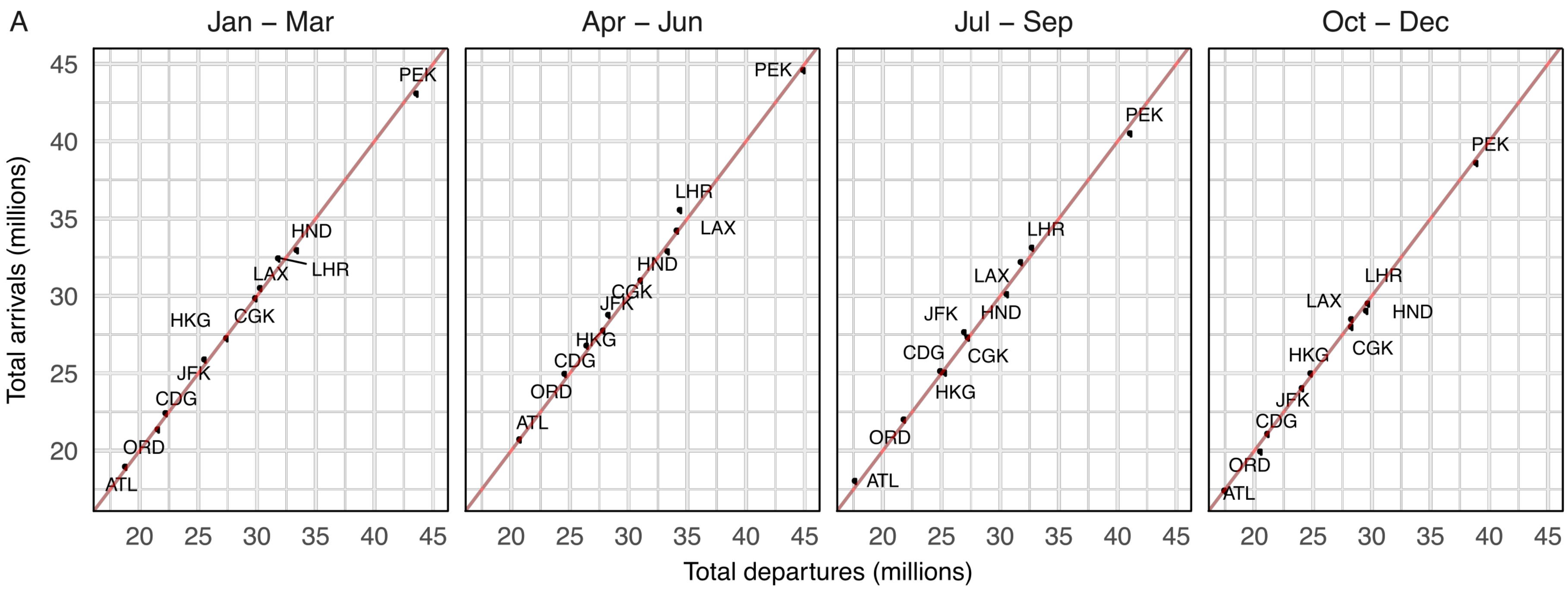


Figure 3.9: Scatter plots representing the number of departing and arriving bookings from each airport globally (log scale) (A); total number of bookings against the difference between departing and arriving bookings (net difference)(B); and total number of connecting bookings against the total difference $((B_{arr} + B_{dep}) - B_{con})$ of bookings (C). Airports are represented by their three letter IATA code.
Note: the red lines represent the line of equality (x=y) (A) and the line of no difference (y=0) (B and C).

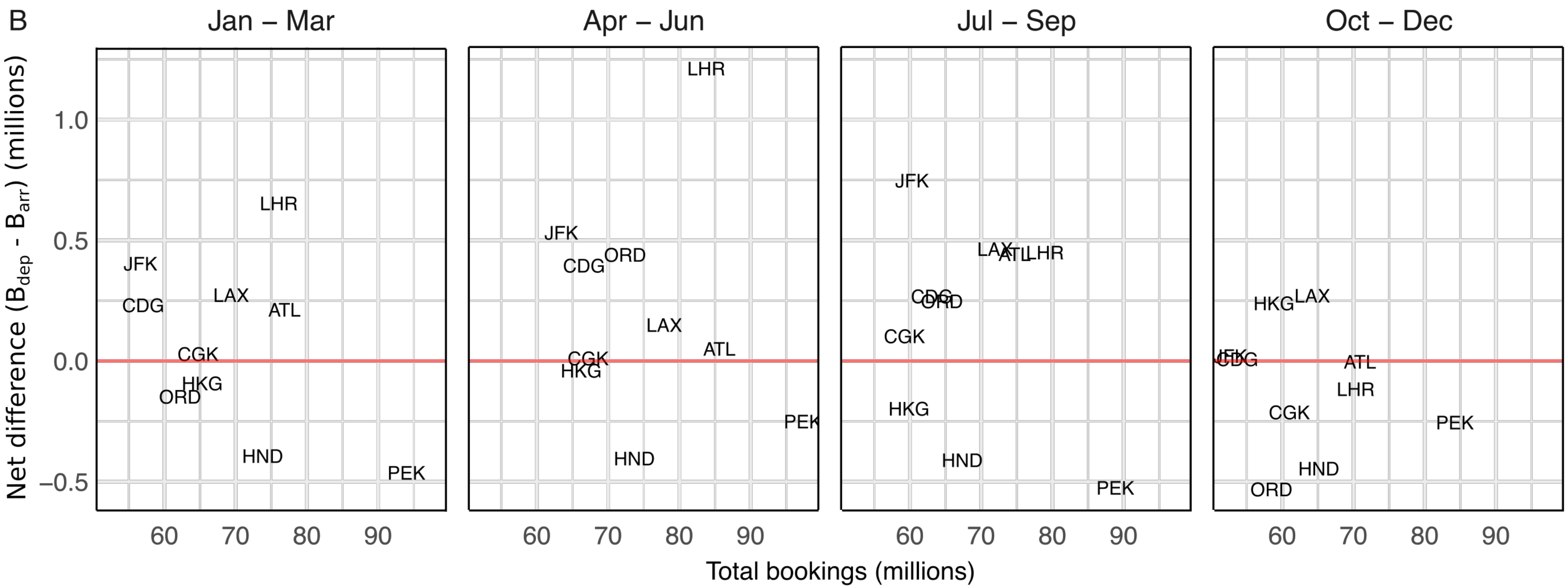
A more detailed analysis of the ten busiest airports globally (**Table 3.1**) was conducted, to determine whether any seasonal variations could be seen (**Figure 3.10**). The ten busiest airports were defined as those with the highest cumulative number of bookings (departing, connecting and arriving) throughout the 64 months. Along with the number of bookings, the percentage of global bookings traveling through these airports is shown in **Table 3.1**.

From **Figure 3.10 A**, a seasonal trend can be seen, where all airports had higher cumulative total bookings between April and June rather than between July and September, at odds with the global seasonal trend. The larger number of bookings in the second quarter may reflect variations in airline routes available between these months. As described by Mao *et al.* (2015), a larger number of air routes may be available from smaller airports, allowing passengers to avoid larger airports over the third quarter. The airports also showed the smallest number of bookings between October and December, which is in agreement with the annual trends seen before in the OAG data. When looking at airports acting as sinks and sources (**Figure 3.10 B**), different patterns arose depending on the airport, such as LHR having the largest number of B_{arr} between April and June but had most B_{dep} between October and December. JFK showed a steady increase in number of bookings from the fourth quarter, peaking in July to September. However, LAX showed a peak in July to September, following a trough in April to June. This example shows that even among the busiest airports, the airline patterns can vary significantly. From this selection of airports, the largest number of B_{con} (**Figure 3.10 C**) was seen in ATL throughout the year. PEK showed the largest number of B_{con} between April and June (7.78 million bookings) and the least in the final quarter of the year (6.69 million bookings). Overall, all but one airport (HKG) had the largest number of B_{con} in the second quarter of the year, whereas HKG had most B_{con} in the first quarter. Finally, all airports, except CGK, had fewest B_{con} during the last quarter of the year, reflecting the limited air routes available as well as the fewest number of bookings.



(Figure 3.10 continues on next page)

(Figure3.10 continued)



(Figure 3.10 continues on next page)

(Figure 3.10 continued)

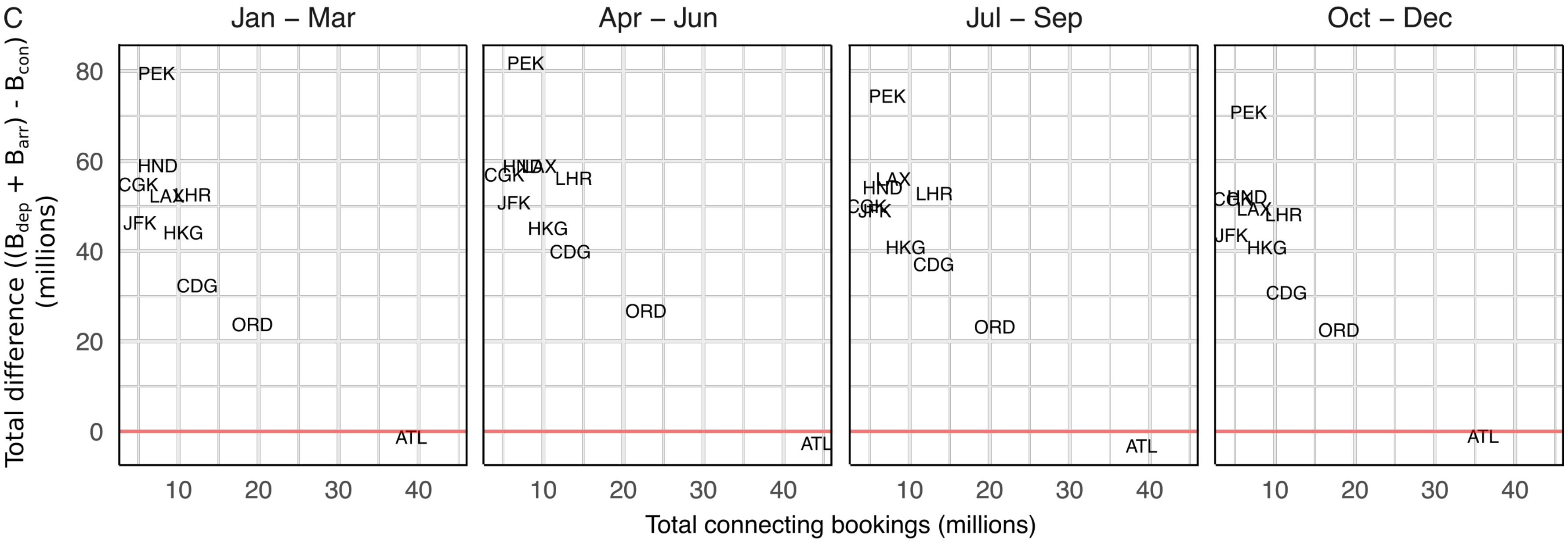


Figure 3.10: Scatter plots representing, for the ten busiest airports globally, the quarterly distribution of departing and arriving OAG bookings (A); quarterly bookings against net difference (B_{arr} - B_{dep}) of bookings (B); and the total number of connecting bookings against the total difference ((B_{arr} + B_{dep}) - B_{con}) of bookings (C).
Note: the red lines represent the line of equality (x=y) (A) and line of no difference (y=0) (B and C).

Data limitations and biases

Dealing with stations

There was a total of 6,726 airport location codes recorded in the OAG data base, of which 669 (9.9%) were railway, bus stations or ferry terminals, with their own IATA code. The presence of railway and bus stations in the data was already noticed by Bobashev *et al.* (2008). There may be additional stations whose 'Airport.Name' was not recorded as a station as such (*i.e.* did not explicitly mention any combination of the terms 'rail', 'railway', 'station', 'stop', 'service', 'ferry' or 'port'), and would therefore be missed in this selection. These stations were found across 31 country names, and consisted of 562 railway stations, 101 bus stations (stops and services also included) and six ferry ports. Canada, Norway and Spain recorded the highest number of these stations with 82, 80 and 71, respectively; and the United Kingdom recorded 62 of them (see **Table 3.2**).

It was ultimately decided to keep these stations within the data itself as the number of bookings associated with them was small relative to the global number of bookings. Additionally, some bookings used stations in their transfer routings, and with no clear airport nearby to transfer the bookings to, this added a level of complexity as there was no clear way of inferring routes via these stations.

UK airports

The complementary data to OAG data set, listing all airport codes and accompanying country names, recorded 167 airports in the UK. However, 61 were railway stations, bus stops or ferry terminals (**Table 3.2**). From the final 106 unique airport names, three airports had duplicated IATA codes: Bristol (three codes in total), Cardiff (two codes) and Exeter (two codes). However, each of these airports only had one valid IATA code attached to routings. The locations of UK airports present in the data is shown in **Figure 3.11**.

Table 3.2: List of railway and bus stations, with accompanying IATA codes and full names (according to OAG), found in the data for the United Kingdom only.

IATA code	Airport name	IATA code	Airport name	IATA code	Airport name
BSH	Brighton Rail Station	TTY	Taunton Bus Station	XVB	Stafford Rail Station
CHW	Cheltenham Bus Station	USX	St Austell Rail Station	XVC	Crewe Rail Station
CLB	Colchester Bus Station	WNC	Winchester Bus Station	XVG	Darlington Rail Station
OXQ	Oxford Rail Station	XGM	Grantham Rail Station	XVH	Peterborough Rail Station
PCW	Par Rail Station	XNE	Newport Rail Station	XVJ	Stevenage Rail Station
POQ	Poole Bus Station	XNK	Newark Northgate Rail Station	XVU	Durham Rail Station
QDH	Ashford International Rail Station	XNM	Nottingham Rail Station	XVW	Wolverhampton Rail Station
QEW	Leicester Rail Station	XNO	Northallerton Rail Station	XWD	Wakefield Westgate Rail Station
QQD	Dover Rail Station	XNV	Nuneaton Rail Station	XWH	Stoke On Trent Rail Station
QQH	Harwich Rail Station	XPF	Penrith Rail Station	XWN	Warrington Bank Quay Rail Station
QQK	London Kings Cross Rail Station	XPT	Preston Rail Station	XWS	Swindon Rail Station
QQM	Manchester Piccadilly Rail Station	XQE	London Ebbsfleet International Rail Station	ZDU	Dundee Rail Station
QQN	Birmingham New Street Rail Station	XQG	Berwick-upon-Tweed Rail Station	ZEP	London Victoria Rail Station
QQP	London Paddington Rail Station	XQH	Derby Rail Station	ZFI	Chesterfield Bus Station
QQR	Ramsgate Rail Station	XQL	Lancaster Rail Station	ZGB	Nottingham Bus Station
QQS	London St Pancras International Rail Station	XRC	Runcorn Rail Station	ZGG	Glasgow Central Rail Station
QQU	London Euston Rail Station	XRE	Reading Rail Station	ZIV	Inverness Rail Station
QQW	London Waterloo Rail Station	XRU	Rugby Rail Station	ZLS	London Liverpool Street Rail Station
QQX	Bath Spa Rail Station	XSR	Salisbury Rail Station	ZXA	Aberdeen Rail Station
QQY	York Rail Station	XTK	Thirsk Rail Station	ZXE	Edinburgh Rail Station
RNW	Ringwood Bus Station	XVA	Stockport Rail Station		

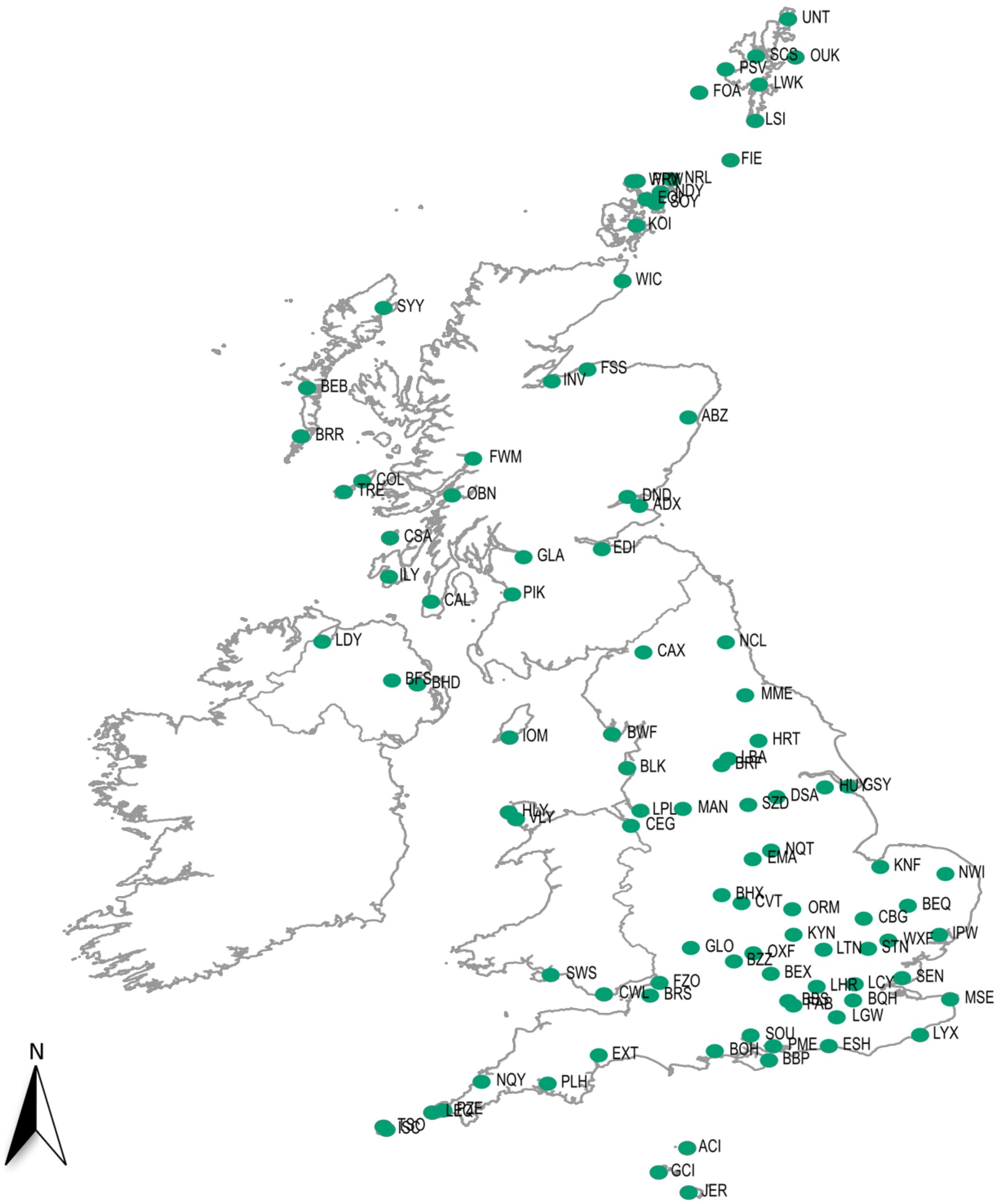


Figure 3.11: Map of airport locations used in OAG with their IATA codes, in the United Kingdom, including Isle of Man, Guernsey and Jersey.

Increased connectivity

When working with the data for the global analysis (chapter 6), it became apparent that in 2014 there was a sudden increase in number of airport connections (degree) present in the data (**Figure 3.12**). This was true for both the in degree (number of direct connections on inbound leg) and out degree (number of direct connections on departing leg). This is likely to be artificial, as such a large increase over a short period of time (one month) is unlikely to be caused by a large amount of additional airports or bookings being included in the data. It is likely that this rise in connectivity may be a result of a change in data collection methods. However, this was neither confirmed by the company, or detailed in the collection methods provided by them (OAG, 2017). Any analysis done using this data may be partially wrong given that the period prior to 2014 may be an underestimate of the total number of bookings, or that the period from 2014 onward may be an overestimate of the data. No adjustments could be made accordingly as it was not clear which were the correct values to use (pre or post 2014).

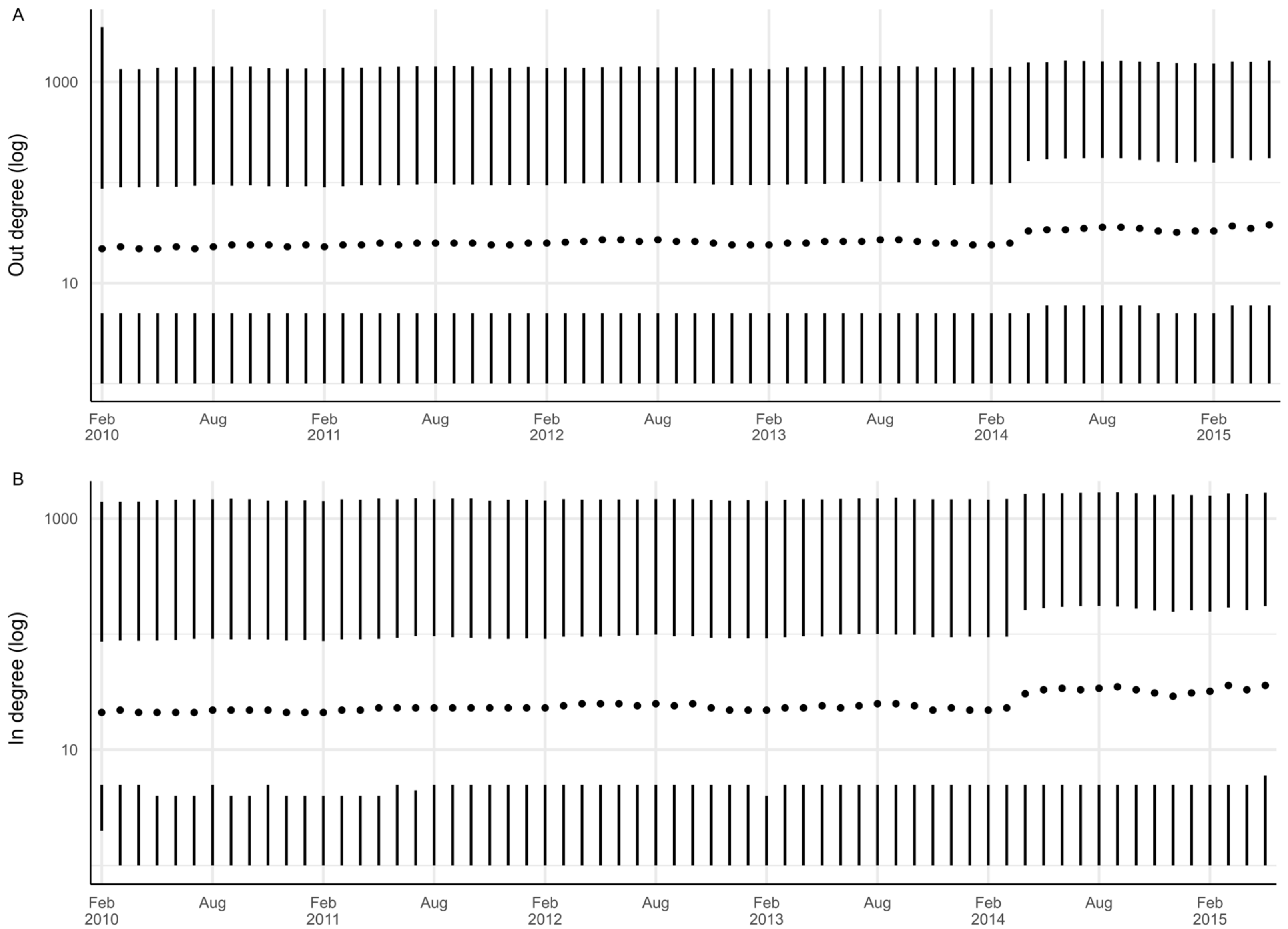


Figure 3.12: Tufte style boxplots of global airports out (A) and in (B) degrees, by month and year, on a log scale.
Note: the points represent the mean, and the vertical lines represent the whiskers.

Discussion

The author described and presented the OAG Traffic Analyser data set, using a number of methods with the intention of validating the internal consistency of the data and identifying missing information and trends. Although several aspects of the collection methods were unclear, the data showed clear seasonal patterns with the most bookings recorded between July and September every year.

OAG data allowed for an initial understanding of the global airline network, as it was possible to determine seasonal trends of the ten busiest airports, but there was no information on passenger demographics. From the Traffic Analyser data set, it was clear that the number of connecting bookings (B_{con}) globally was much smaller than the number of departing (B_{dep}) and arriving (B_{arr}) bookings, but that each airport had its own associated travel patterns and played its own role in the global network. For example, airports like ATL and CLT were primarily used for connections whereas PEK had very few connections but a large number of departing and arriving bookings. The ten busiest airports showed seasonal trends that were not in agreement with the global average annual trends, such as having the largest number of bookings between April to June rather than July to September, and the fewest bookings in October to December rather than January to March. This may be an artefact of the network: as the number of passengers increases, so does the number of air routes available, allowing passengers to avoid busier airports (Kraft and Havlíková, 2016).

Although the data set gave some understanding of the airline network, a number of issues were encountered. First of all, the data included railway and bus stations as well as ferry terminals, with their own IATA codes. Given their locations within routings (as connecting points) and that it was not possible to assign these bookings to an actual airport code, these stations were left in the routings. The company's data description (OAG, 2017; OAG, 2016c) provided little information on their data collection methods and did not state the inclusion of these stations as part of their data.

Additionally, through exploration of airport connectivity it became apparent that there was an increase in connectivity from 2014 onward. The cause of this change was likely to be a change in methodology for the company's data collection; however, this was not confirmed by the company or their data collection method descriptions. Such an important change in connectivity is likely to impact any analysis done using this data. This is another example of the importance of understanding and describing one's data to determine whether it is fit for the task required.

As with all data sets, having a good understanding of one's data set before any analysis is done, is important to ensure the correct results are reached (Emanuelson and Egenvall, 2014). As third-party data were not collected for the same specific research question, researchers have no or little control over its collection methods and quality, and therefore need to critically assess its usefulness prior to use. This evaluation of usefulness needs to be determined for the research project in question (Emanuelson and Egenvall, 2014).

Although the OAG data is historic at the time of writing, the use of such data may provide insights to understand the future spread of outbreaks through air travel.

Data dictionary

Variable name	Definition	Values	Units	Format	Source
Arr.Airport.Cd	IATA code of arriving airport.	Too many to list here*.	N/A	Character	OAG download
Arr.Country.Name	Arriving Country Name, country where the journey (not country) ends.	See Appendix for table of OAG country names.	N/A	Character	Generated from data
Arr.Region.Name	Arriving Region Name, region where the journey (not country) ends.	See Appendix for table of OAG region names.	N/A	Character	Generated from data
Bookings.Adjusted.	Number of adjusted bookings, between airports, including indirect bookings.	Ranging from 0 to several thousand.	Bookings	Integer	OAG download
Bookings.Unadjusted.	Number of unadjusted bookings.	Ranging from 0 to several thousand.	Bookings	Integer	OAG download
Dep.Airport.Cd	IATA code of departing airport.	Too many to list here.	N/A	Character	OAG download
Dep.Country.Name	Departing Country Name, country where the journey (not country) begins.	See Appendix for table of OAG country names.	N/A	Character	Generated from data
Dep.Region.Name	Departing Region Name, region where the journey (not country) begins.	See Appendix for table of OAG region names.	N/A	Character	Generated from data
Distance...Km	Distance from departing to arriving airport in kilometers.	Too many to list here.	Kilometers	Integer	OAG download
Journey_type	Leg of journey to which the booking refers to.	'Outbound international', 'Inbound international', 'International domestic', 'Domestic', 'International other' and 'Other'.	N/A	Character	Generated from data
Number_connections	Number of connections present in a given routing.	From 0 to 2	Connections	Integer	Generated from data
Point.of.Origin.Cd	IATA code for the airport from which the journey begins.	Too many to list here	N/A	Character	OAG download
Point.of.Sale...Country	Country from which the booking was purchased.	See Appendix for table of OAG region names.	N/A	Character	OAG download
Routing	List of departing, arriving and connecting airports visited by a given bookings, given by their IATA code.	Too many to list here	N/A	Character	OAG download
TimeSeries	Time series for a given booking, in month and years.	Years: '2010' to '2015' Months: 'January' to 'December'.	Time	Character	OAG download

* IATA codes can be found at: www.iata.org/publications/Pages/code-search.aspx

Chapter 4 – Airline data validation

Preamble

As well as having a good understanding of the data used, researchers should be encouraged to validate third hand data against independent and comparable data sets. As far as the author is aware, such an analysis has not yet been reported, and therefore unlikely to have been undertaken. This analysis allowed for an understanding of what OAG means when using the term ‘bookings’ and whether this differs from passenger numbers.

Abstract

Mathematical modellers are known to use commercial airline data sets in models that help inform governmental policies. However, their reporting has previously been shown to be insufficient for replication by other research groups, and is understood not to have been validated to determine whether any biases or anomalies are present within. A detailed description of the data set was undertaken in the previous chapter, and the aim of this chapter was to validate the OAG data against independent data sets and determine how many passengers were considered for each OAG booking.

Four independent yet comparable and open access data sets were acquired and described in depth before being used to validate the OAG data. These were from the US Department of Transport (USDoT), Port Authorities of New York and New Jersey (PANYNJ), Civil Aviation Authority (CAA) and the Office for National Statistics. Several comparisons were done, including temporal, ratios and mixed effects regression analyses.

All data sets showed the same seasonal patterns, with most travel seen between July and September, and the least seen between January and March. When directly comparing the data sets against each other, the TravelPac and USDoT matched closely with OAG, whereas the CAA and PANYNJ did not match as well. When determining the number of passengers considered per booking, the TravelPac and USDoT both showed a value of around one passenger per booking, whereas the CAA and PANYNJ showed values ranging between one and three passengers per booking and above.

With each data set used for any analysis, it is important for researchers to have a clear understanding of what the data represents, especially for expensive data and/or if the methods are not clearly detailed and easily available. Although there are a wide variety of data sets available to modellers, different data sets will represent airline travel in a different way that may not be what the researchers need or should use. Therefore, researchers should be encouraged to not only report the data they are using to allow accurate reproduction of their work, but they should also be encouraged to validate it, at least in part, against independent yet comparable data sets.

Introduction

Researchers must know what the data they are intending to use represents, and what it contains, to avoid erroneous conclusions. Therefore, a comparison (or validation) of the whole or partial data set against another source should be done. However, complications may arise when undertaking this task and may only provide a partial degree of completeness and agreement between both data sets, especially if neither is thought to be a gold standard. Secondary data needs to be correct, however, the level of completeness varies with context (Emanuelson and Egenvall, 2014).

As was highlighted in the systematic review, validation of the airline data used against a comparable and independent data set is not frequently performed. A large number of data sources (n=36 identified) have been used by research groups to build mathematical models that may in turn be used to inform policy makers about potential importation risks. Some of these sources are expensive to access and marketed as highly accurate (International Air Transport Association (IATA) and OAG data services, for example), but their collection methods are unclear due to commercial sensitivities. Open access data sets are available (US Department of Transport, UK Civil Aviation Authorities, for example) but have limitations. For example, data on passengers published by governments will be geographically restrained to that particular country. Furthermore, temporal resolutions vary between sources, such as monthly, quarterly or annually. Finally, data sources often have different and incompatible variables.

There is currently no single data source representing airline passenger traffic that has been agreed upon by the mathematical modelling community as being the most representative or best to use. However, the implications of using different data sources can be significant; both for estimating the risk of disease importation and for correct policy planning. For example, during the West African Ebola (EBV) outbreak of 2013-15, the United States of America was not estimated as a country with a high risk of EBV importation according to some research groups (Bogoch *et al.*, 2015). However, the USA saw the highest number of imported cases outside of West Africa (Elmahdawy *et al.*, 2017). The data used for these analyses only considered direct flights between the West African countries (Liberia, Sierra Leone and Guinea) and the United States (Bogoch *et al.*, 2015). Given that these were suspended during the outbreak (European Centre for Disease Prevention and Control, 2014), using indirect flights (routings with at least one stop) would have given a more accurate representation of the risk of importing Ebola cases. This example highlights the need for consistency between research groups in terms of which data sets to use by the modelling community, and the importance of using accurate information. Additionally, researchers should make a clear description of the data type they are using clear to the reader (whether this is direct or indirect flights, passenger number or aircraft capacity estimates) as using different data types may lead to different conclusions from the models.

This chapter aimed to compare open access data sets with the closed access and expensive Traffic Analyser data set downloaded from OAG between August 2014 and July 2015 (thereafter referred to as OAG). This data set was described in the previous chapter, and will be directly compared against four independent and open access data sets, from various sources. This comparison serves as a data validation of the OAG data, and can be considered a first step in determining a commonly agreed data set for future work in infectious disease modelling.

Methods

Describing open access data sets

To assess how the OAG data compares to open access and independent measures of passenger traffic, it was directly compared to four open access data sets, described in **Table 4.1**. Two of these data sets (Port Authorities of New York and New Jersey (PANYNJ) and the United States Department of Transport (USDoT)) represent United States airline passenger information, with varying levels of detail, whereas the Civil Aviation Authorities (CAA) and Office for National Statistics represent airline patterns relevant to the United Kingdom (UK). These two countries were chosen as the United States represents a large portion (22% of global passengers carried) of global airline traffic (The World Bank, 2017) and the UK is central to this thesis. The full OAG data was stratified and/or aggregated appropriately to permit a direct comparison with each data set in turn (**Table 4.1**).

Table 4.1: Decription of the data sets used for OAG data validation.

Data provider[1]	Port Authority of New York and New Jersey (PANYNJ)	United States Department of Transport	Office for National Statistics (UK)	Civil Aviation Authorities (UK)
Data set name	"Monthly Summaries of airport activities"	"Origin and Destination survey: DB1B Market"	"TravelPac"	"International Air Passenger Route Analysis"
Date downloaded	May 2016	May 2017	August 2015	January 2017
Date-range covered	January 2010 to December 2015	April 2010 to March 2015[2]	April 2010 to March 2015[2]	February 2010 to May 2015
Temporal resolution	Monthly	Quarterly	Quarterly	Monthly
Geographic resolution	Airport	Airport	Country	Airport and country
Data nature (according to source)	Passenger numbers; domestic and international; departure airport only.	Passenger numbers, domestic only; origin and destinations included.	Passenger visits, international.	Passenger numbers to and from reporting airport pairs.
Data interpretation (according to authors)	Passenger numbers – international and domestic flights, departing from specified airports.	Passenger numbers within the USA and territories (Puerto Rico, Guam, Virgin Islands, Mariana Islands and American Samoa[3]).	Passenger numbers (UK residents departing and overseas residents arriving) according to country visited (UK residents) or of origin (overseas residents).	Passenger numbers – return trips – between airport pairs.
Collection method	Unknown	"Origin and destination survey"	"International Passenger Survey"	"CAA Passenger survey reports"
Corresponding OAG data subset	Passengers departing specified airports internationally or domestically; unknown trip point of origin.	Booking numbers for routings departing and arriving within the USA and associated territories; unknown trip point of origin.	UK residents: Trip originated in UK, passengers returning into UK. Overseas residents: Trip originated internationally, passengers departing from UK.	Passengers leaving the UK and returning from abroad, combined; trip point of origin unknown.
Comments	One data base divided into 2: international and domestic.	The data represents a 10% sample of passengers.	Can differentiate between UK and overseas residents.	No journey leg, therefore flow direction unknown.

[1] PANYNJ: www.panynj.gov/airports/traffic-statistics.html
USDoT:
www.transtats.bts.gov/Tables.asp?DB_ID=125&DB_Name=Airline%20Origin%20and%20Destination%20Survey%20%28DB1B%29&DB_Short_Name=Origin%20and%20Destination%20Survey
ONS: www.ons.gov.uk/peoplepopulationandcommunity/leisureandtourism/datasets/travelpac
CAA: www.caa.co.uk/Data-and-analysis/UK-aviation-market/Airports/Datasets/UK-airport-data/

[2] January 2010 and June 2015 were not available in the OAG data; therefore a full comparison of these two quarters (quarter 1 2010 and quarter 2 2015) was not possible and download of data done accordingly.

[3] Guam and the US Virgin Islands are "territories of the United States"; Puerto Rico is a "territory of the US with commonwealth status"; the Mariana Islands are a "commonwealth in political union with the US"; the American Samoan islands are "self-governing territory of the US" (Central Intelligence Agency, 2013).

Port Authorities of New York and New Jersey (PANYNJ)

Two data sets were downloaded from the PANYNJ website: 'monthly domestic' and 'monthly international' passenger numbers, leaving from specific airports, but with no defined destination. International passenger numbers were available from three airports: John F Kennedy International (JFK), La Guardia (LGA) and Newark Liberty International (EWR) Airports. The domestic passenger data also included a fourth airport, Stewart International (SWF) Airport. No detail was given regarding the passenger's final destination, any stops in their journey or where their trip originates from (point of origin), which may be different to the departing airport.

Therefore, the equivalent OAG subset was selected as bookings departing from the same airports (JFK, LGA, EWR and SWF), and the arrival country, categorised as 'USA' or other. If the arrival country was selected as 'USA', the data were used for the domestic data comparison; if the destination country was not the USA, the data were used for the international data comparison.

United States Department of Transport (USDoT)

The data were downloaded from the "Origin and destination survey: Market" database, containing passenger numbers between a specified origin and destination, at a given annual quarter. This data represented a 10% sample of the passengers interviewed at airports, but does not mention the number of interviewed passengers. The geographical extent of the data included the United States as well as some overseas territories (**Table 4.1**), however, no stop over or point of origin airports were recorded. The original data represented a 10% sample of total passengers in the USA and territories, and were inflated by a factor of ten to correspond to the OAG booking numbers.

The equivalent OAG data subset was selected as bookings with routings within the USA and territories, without specifying a point of origin. The monthly data was then aggregated to match the quarterly time resolution of the American data set.

Office for National Statistics (ONS, UK)

The Office for National Statistics (ONS) collects, analyses and publishes nation-wide statistics about the United Kingdom's economy and population (Office for National Statistics, 2017). The International Passenger Survey (IPS) has been generated by the ONS since 1961, from which the quarterly data is derived resulting in the TravelPac data set. The IPS data is derived from ongoing passenger surveys collected through face-to-face interviews with voluntary participants at various ports of entry into the UK, by various modes of transport: air, sea or road. The surveys only represent a small proportion of travellers (0.2% in 2009) travelling through specific airports (with more than 1 million passengers per year) therefore the numbers are inflated (by the ONS) using weightings to give national estimates and generate the TravelPac data set (Office for National Statistics, 2014).

For the purposes of this analysis the data were first restricted to passengers travelling through airports only ('air'), and a first comparison of the data was done without considering visitor's country of residence. A second analysis was then done by separating UK from overseas residents.

The OAG data was aggregated to a quarterly time series to match that of TravelPac. Both equivalent data sets (UK or overseas residents) were then compared for all full quarters available: Q2 2010 to Q1 2015. The OAG equivalent subset also separated into UK and overseas residents. Bookings representing UK residents were selected as routings with a point of origin in the UK, and the UK as the arriving country name, but an international departing country name. In contrast, bookings representing overseas residents were routings with point of origin anywhere outside the UK and with a departing country name other than the UK, but the UK as the destination country.

Civil Aviation Authorities (CAA)

This data set represents the monthly number of passengers who travelled between airport pairs (one of which in the UK, the other international), between February 2010 and May 2015. This number of passengers represented the flow of returning passengers between specified airport pairs and included indirect flights. However, the direction of passenger flow between these airport pairs was not specified. It was unclear how many passengers travelled from LHR to JFK and from JFK to LHR.

The corresponding OAG data was selected by sub-setting routings departing from the UK and arriving internationally, as well as bookings arriving in the UK but departing from the same international airports, both with unknown points of origin. The OAG point of origin could not be specified as it was unclear from the CAA data whether or not these passengers were UK residents. These bookings were aggregated by month and collated. The airport names used in these data did not match those of the OAG data and no IATA code was assigned to them. Airport names were first cleaned to match those of OAG. If some CAA airport names were not specific enough (*e.g*. “oil rigs”) or no OAG equivalent could be found (*e.g.* “Lyon(Bron)”) the passenger counts for the associated airport name were excluded from the data.

Comparison and interpretation of independent and OAG data sets

The OAG data were compared to several open access data sets in several ways. First, a temporal pattern comparison was undertaken at the finest temporal resolution available, monthly or quarterly. The total aggregated number of passengers per bookings per month or quarter was plotted against time. The PANYNJ data had the smallest geographic resolution, with four airports within 62 miles (100 km) of each other, therefore all airports time variations were included in the comparison.

After aggregating the number of passenger or visits and bookings to the finest geographic resolution (airport or country) and not taking the temporal resolution into account, a direct comparison of the passenger flow was done. This calculation of passenger (or visits) per bookings ratio gave a first understanding of the number of passenger (or visitors) considered for each booking per airport or country.

These passenger-per-booking ratios were then aggregated by month or quarter and plotted against time to determine any seasonal trends. Given the large amount of noise present in the data, the ratios for USDoT, TravelPac and CAA were restricted to values below 7 (USDoT and TravelPac) or 10 (CAA), and further restricted to only include specified airports.

Finally, these aggregated ratios were plotted by month or quarter and separated by year. This allowed to determine any seasonal trends within each ratios and see whether the ratios are consistent across the year.

For all data sets, except PANYNJ, a selection of countries and airports was also plotted for extra clarity. These were chosen as the ten busiest UK airports (CAA and USDoT), the ten countries most visited by UK residents or from which residents arrived from (TravelPac).

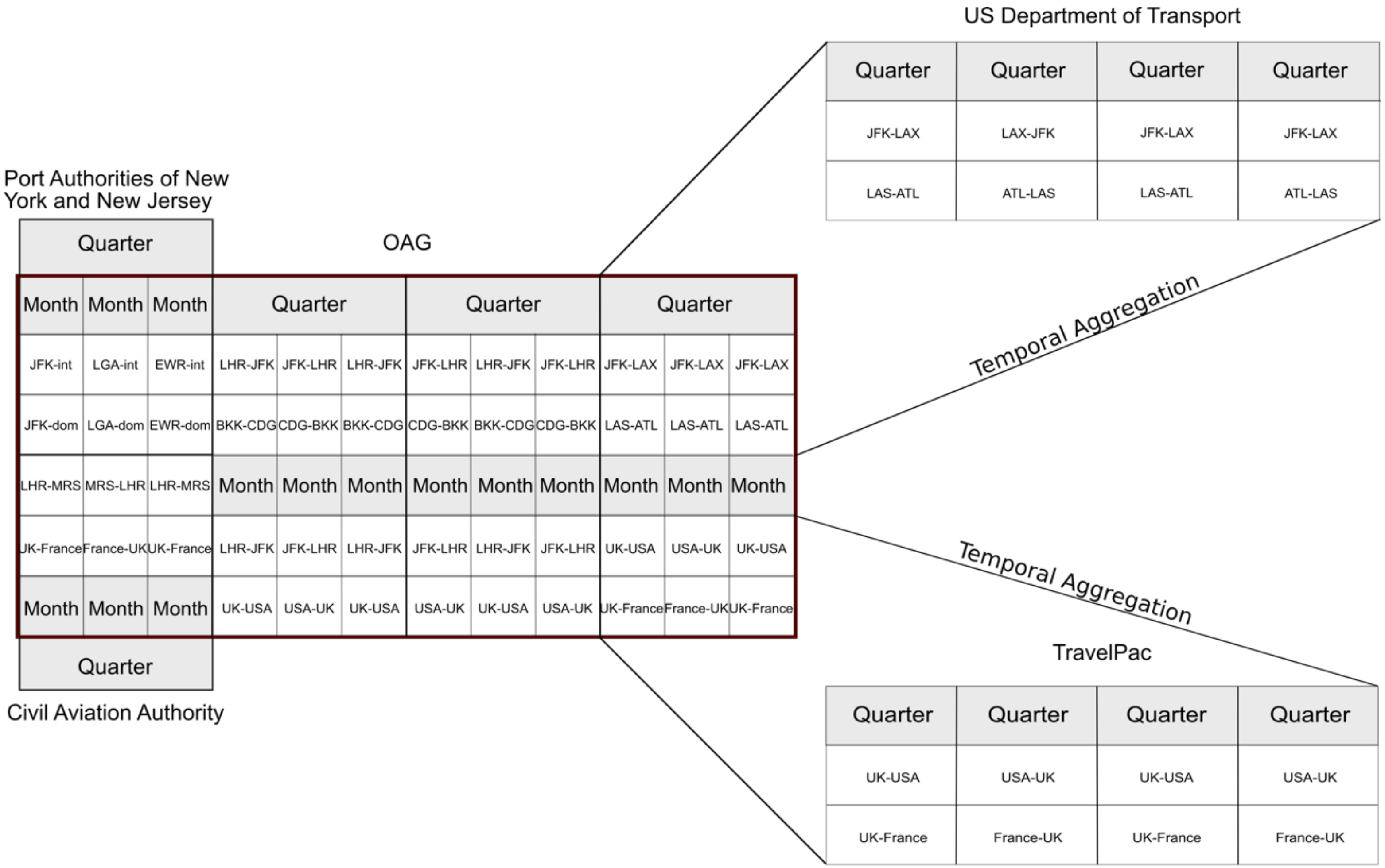


Figure 4.1: Visual representation of each data set used for the validation in comparison with OAG.
Note: the routings are for illustrative purposes only.

Regression analysis

A random effects negative binomial model was developed to determine the average number of passengers (from open access data) per OAG booking and whether this varied between geographic resolution (airport or country level). It was assumed that the passenger number departing from each airport (y_i) was drawn from a negative binomial distribution (with mean μ_i and dispersal parameter r), and the relationship modelled with departing passenger bookings (x_i) as a regression model with zero-intercept and a random slope for each departure airport (β_i). The author assumed that the random slopes were normally distributed around zero with standard deviation τ; the slope of the regression could be interpreted as an estimated mean number of passengers per booking. Specifically,

$$y_i \sim NB(\mu_i, r)$$

$$\log(\mu_i) = \beta_i \log(x_i)$$

$$\beta_i \sim N(0, \tau^2).$$

This model was run at both airport and country geographic resolutions, with only positive values used from the airports data set. All analyses were done using the lme4 package in R v3.4.1 (R Core Team, 2017). The 95% confidence intervals were not shown here as the author was only considering the point estimate of each distribution.

Results

Correspondence of temporal patterns

There were clear seasonal patterns in all data sets, visits or passenger numbers with peaks in the months of July and August or Q3, and troughs between the months of November and February or Q1, in agreement with the OAG data (**Figure 4.2**). The monthly data sets (PANYNJ and CAA) showed additional small peaks on or around December. All validation data sets showed larger numbers of passenger visits than their corresponding OAG counter-part, with varying proportions.

For the PANYNJ airports, the overall variation in passenger numbers departing internationally from John F Kennedy International Airport (JFK) was greater than the corresponding number of OAG bookings (**Figure 4.2 A**). During the time period considered, the number of passengers departing (PANYNJ) from EWR closely resembled the number of OAG bookings departing from JFK. Regarding domestic departures (**Figure 4.2 B**), the number of passengers leaving EWR, JFK and La Guardia (LGA) was much larger than the number of OAG bookings leaving these same airports. Although the number of flights and bookings for EWR were on a much smaller scale, similar trends to the other airports could be seen.

The USDoT (**Figure 4.2 C**) recorded more passengers within the USA and territories, than OAG bookings, with 2,224.2 million passengers and 2,217.4 million bookings. Interestingly, in the USDoT data comparison, the peaks of passenger and booking numbers were seen in Q2 (April to June) and not in Q3 (July to September) as was the case in all other data sets.

The TravelPac data set (**Figure 4.2 D**) also showed larger numbers of visits than OAG bookings, with important inter-seasonal variations in the number of visits and bookings recorded (max visits= 23.8 million (2014 Q3), max bookings= 11.4 million (2014 Q3)).

With regards to the CAA data set (**Figure 4.2 E**), the total passenger numbers between airport pairs was much larger than the equivalent OAG number of bookings. Additional smaller peaks were seen in the number of passengers in December of every year but were not discernible in the equivalent OAG bookings.

Direct comparison of passenger/visitor numbers against OAG bookings by geographic resolution

OAG bookings were compared with the passenger or visit numbers from each other data set. Generally, similar patterns of OAG underestimation were evident in these comparisons, except for TravelPac and USDoT where the linear trends were in agreement with the OAG bookings (**Figure 4.3**).

For the PANYNJ data, more passenger numbers (136.1 million) were reported than OAG bookings (59.0 million) and specific patterns were seen (**Figure 4.3 A** and **B**). For example, JFK was the main airport used for departing international flights, whereas EWR, JFK and LGA saw overlapping number of domestic passengers and bookings leaving each airport. Additionally, for both travel directions, the PANYNJ passenger numbers were larger than the corresponding OAG bookings for JFK and EWR. However, there was more agreement between the data sets regarding passengers departing internationally from La Guardia and Stewart International (SWF) airports.

The USDoT data set also shows a good overall relationship between both data sets, with both data sets over and under estimating passenger numbers and bookings for varying time points and airports (**Figure 4.3 C**). The airports with the largest number of bookings and passengers were Los Angeles International (LAX) (82.6 million passengers and 81.3 million bookings), Las Vegas McCarren International (LAS) (74.0 million passengers and 74.6 million bookings) and Orlando International (MCO) (67.3 million passengers and 66.7 million bookings) airports.

The TravelPac data set showed good agreement with its OAG equivalent bookings comparison data (**Figure 4.3 D**). It can be noted that the number of visits and bookings recorded for Spain from both data sets was much larger than those from any other country, with important variations ranging from 1.7 million to 5.1 million visits and 1.7 million and 4.8 million bookings. The next most important countries in terms of visits and bookings were the USA (28.0 million visits and 26.7 million bookings), followed by Germany (19.4 million visits and 23.8 million bookings) and France (21.9 million visits and 19.1 million bookings).

The direct comparison using CAA passenger numbers (**Figure 4.3 E**) showed the largest number of passengers travelled via London Heathrow, followed by London Gatwick and Manchester airports. Once again, the number of passengers recorded in the CAA data was larger than the equivalent OAG bookings, with 101.9 million passengers and 341.1 million bookings for Heathrow, 60.1 million passengers and 155.1 million bookings for London Gatwick and 31.0 million passengers and 90.4 million bookings for Manchester airports.

Passenger to booking ratios against time

When comparing the passenger per bookings ratios against time the previously distinct seasonal trends were not as clear as previously (**Figure 4.4**).

For the international PANYNJ airports (**Figure 4.4 A_1**), JFK was seen to have a passenger per booking ratios ranging between 1.92 and 2.51 passengers per bookings, whereas EWR ranged between 2.32 and 2.99 passengers per booking and LGA between 0.75 and 1.65 passengers per booking. Overall, these ratios showed an increasing trend over time. For domestic flights (**Figure 4.4 A_2**), lower ratios ranging between 1.86 and 2.63 were seen for LGA and 1.68 and 2.11 for SWF, whereas JFK and EWR showed ratios ranging between 2.43 and 3.39, and 2.42 and 4.01, respectively. The increasing trend seen in the international ratios was not always seen in the domestic ratios, with SWF and EWR showing decreasing trends.

When considering the ratios for the other three data sets, seasonal trends were once again unclear, unless considering the busiest (most passengers and bookings) airports or countries. As either data set may have had zero or very few passengers, visits or bookings recorded for specific airports or countries, some ratios were therefore much larger than others (**Figure 4.4 B_1**, **C_1** and **D_1**). The graphs shown here were therefore restricted to ratios smaller than 10 (CAA) or 7 (TravelPac and USDoT) passengers per booking for clarity.

For the USDoT total data set (**Figure 4.4 B_2**), ratio values centred on a value of one passenger per booking, with no clear seasonal trend. However, when considering the ten busiest airports (those with the largest combined number of passengers and bookings), all values ranged between 0.83 and 1.15 passengers per booking. The ten busiest airports selected were: Los Angeles International (LAX), Las Vegas McCarren (LAS), Orlando International (MCO), O'Hare International (ORD), Denver International (DEN), Hartsfield–Jackson Atlanta International (ATL), San Francisco International (SFO), Logan International (BOS), La Guardia International (LGA) and Dallas-Fort Worth International (DFW) airports. A peak can be seen in Q4 of 2011 and a trough in Q1 of 2012. From 2011 onward, the ratios of passengers per booking dropped in Q1 and Q3 of each year. The airport with a consistently high ratio value was DEN (ranging between 1.01 and 1.15), whereas LGA airport had a consistently low range of values (ranging between 0.86 and 1.05), in this selection. Generally, these selected airports showed similar patterns to each other, peaking in the same quarters.

For the TravelPac data set (**Figure 4.4 C_1, C_2**), a large variation was seen for the ratio of visits per booking by country, with the majority of ratio centred around a value of one (mean of 1.48). When looking at the ten most visited countries from the same data set, the ratios range between 0.66 and 1.68 visits per booking. The ten busiest countries (in terms of combined visits and bookings) were France, Germany, Greece, Italy, Netherlands, Poland, Portugal, Republic of Ireland, Spain and USA. Some seasonal trends were seen, especially in visits to France and Greece, where visitors to France peaked in Q1 of every year and visits to Greece peaked in Q3 of every year. When considering visits to Spain, which was the country with largest number of visits (**Figure 4.3**), the ratio ranged between 0.88 and 1.17 visits per booking. Additionally, the ratio values for some countries like Germany are always smaller than 0.90 visits per booking.

When selecting the ten busiest UK airports from CAA (**Figure 4.4 D_2**) (with the largest combined passenger and booking numbers), the results showed very similar ratio values to each other, varying between 1.53 and 3.84 passengers per booking. The ten busiest airports selected were: Heathrow (LHR), Gatwick (LGW), London Stansted (STN), Manchester (MAN), Luton (LTN), Birmingham (BHX), Bristol (BRS), Edinburgh (EDI), Liverpool (LPL) and East Midlands (EMA) airports. LHR was seen to have the largest ratios of this selection (mean=3.35 passengers per booking), closely followed by Manchester (MAN) (mean=2.92 passengers per booking), Birmingham (BHX) (mean=2.83 passengers per booking) and London Gatwick (LGW) (mean=2.57 passengers per booking) airports.

Monthly aggregated passengers per booking ratio per year

Some seasonal trends were apparent when aggregating each data set by month or quarter, peaking in July to August or Q3, except for the USDoT which peaked in Q2 (**Figure 4.5**).

The aggregated ratio values for the PANYNJ data (**Figure 4.5 A**), showed a mean of 2.06 passengers per booking for the international passengers ranging from 1.67 to 2.32 (international), which was smaller than for the domestic ratios (mean =2.47 passengers per booking, ranging from 2.18 to 3.03). The seasonal trend was not clear for both subsets, however both showed lower ratio values for January and February than for the rest of the year. For the international passengers, there was a general increasing trend with each year. When considering the domestic passengers, the ratio values were overall stable from February onward with a few spikes from June onward in certain years.

From the USDoT monthly aggregated data (**Figure 4.5 C**), the seasonal patterns previously seen in **Figure 4.4** were not seen here, with ratios peaking in Q2, except for 2012 (peak in Q4). For every year considered here, there was a trough in Q3, as previously seen. The average number of passengers per booking was 1.02 and ranged from 0.97 to 1.09.

From the TravelPac aggregated ratios (**Figure 4.5 D**), a seasonal trend was also determined from these ratio values, with the highest average ratio values seen between July and September every year, contrary to the two data sets seen so far. This followed a slow rise between January-March and April-June. The average ratio values ranged between 0.96 and 1.21 with an average of 1.06.

A clear seasonal trend of monthly ratio values could be seen in the CAA data (**Figure 4.5 E**) with the majority of peaks between May and August every year. A sharp rise and drop in average ratio values was seen on either side of these months with an unusual drop in April 2010. The average ratio values ranged between 1.10 and 1.22 passengers per booking with an average of 1.16.

TravelPac and USDoT data don't ask/consider whether passengers are travelling alone or with at least one other passenger, giving the understanding that there is one passenger per booking.

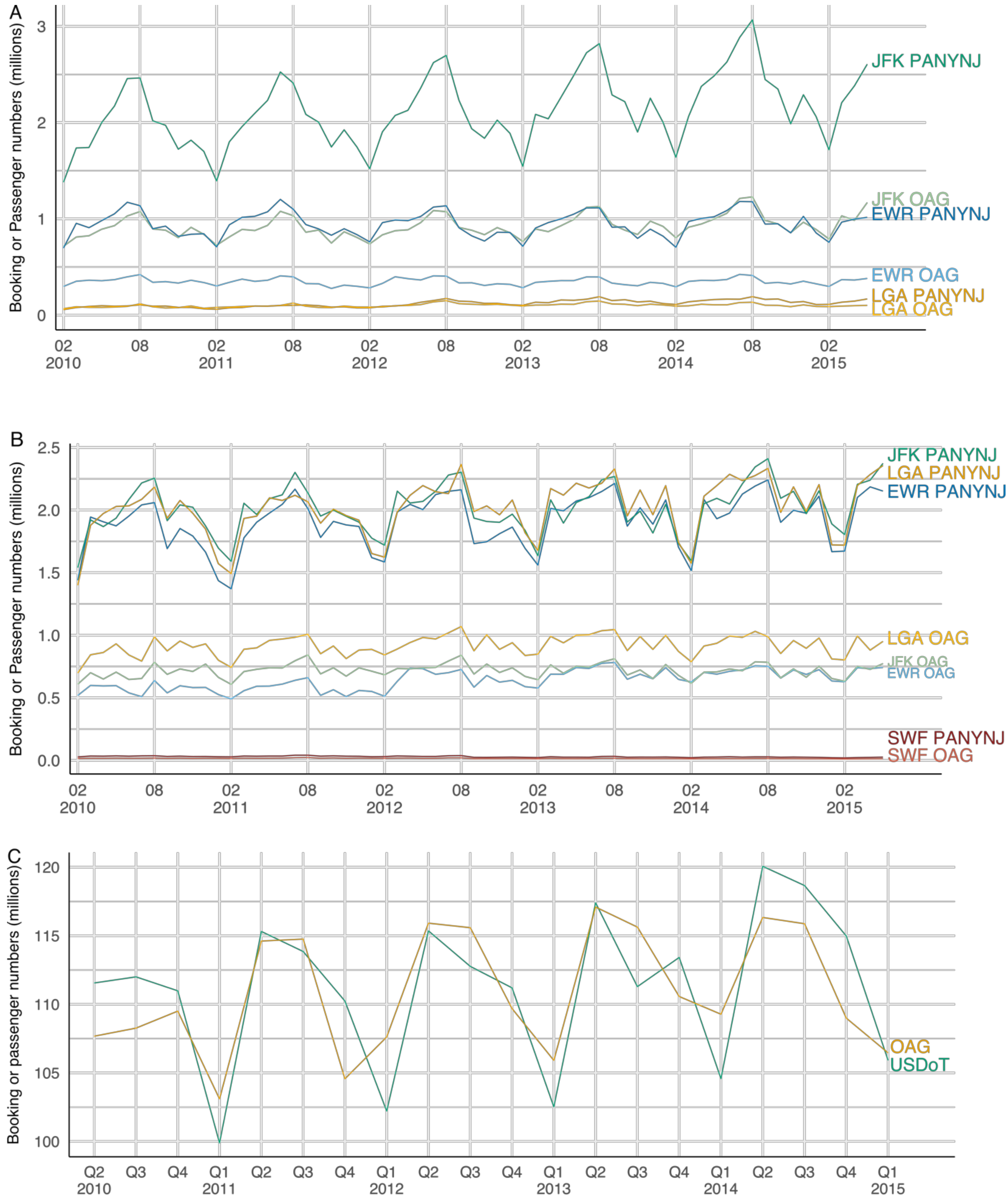
Regression results

When aggregating the data by geography, the negative binomial model showed a median random slope coefficient for all airports of 0.90 and centred around one passenger per booking, with an average of 6.73 (**Figure 4.7** and **Table 4.2**). With regards to the country level model, the median slope coefficient was 0.82, with an average of 1.88. The average values for both data sets were heavily influenced by a number of outliers which had very large slope coefficients, such as the Concord airport, North Carolina in the United States (IATA code 'USA') with a slope coefficient value of 2,913.30 (represented as the point with the highest slope coefficient value in **Figure 4.7A**) and Greenland with a slope coefficient value of 54.36. Additionally, the airport data was heavily influenced by the USDoT data as this represented the majority of the observations (77.0%), whereas the CAA and PANYNJ data included fewer observations (19.0% and 3.9%, respectively).

To better understand the range of values represented in **Figure 4.7**, the range of slope coefficient values were divided into quartiles and colour coded accordingly, prior to their associated airport code mapped (**Figure 4.8**). From these results, it was apparent that when considering airport data (**Figure 4.7A**), the slope coefficient values were centred around one passenger per booking, with variations seen when considering smaller airports, represented by the first and fourth quartile groups in **Figure 4.8**. However, larger airports such as Los Angeles International and Atlanta showed slope coefficients values closer one passenger per booking, with 0.91 and 0.90 passengers per booking respectively. The variations in slope coefficients were likely influenced by the difference in number of passengers and bookings recorded in the data sets. For example, for the Concord airport, the USDoT recorded 18,130 passengers whereas only three were recorded in the equivalent OAG data. Regarding Greenland, the OAG data underestimated the number of bookings to 11, whereas TravelPac recorded 1,123.07 visits.

The collection methods for each data set is likely to influence the slope coefficients, as the data from the Port Authorities of New York and New Jersey seemed to show the least dispersion (mean and median passenger per booking are both equal to 1.04) with values ranging between 0.53 and 1.36 (**Table 4.2**), even with the smallest number of observations. However, airports recorded in this data show slope coefficients centred around 2 passengers per booking, in contrast with data from the US Department of Transport for the same airports.

On the other hand, data set like the TravelPac and US Department of Transport that have been collected through passenger interviews (10% of travellers are interviewed) and have been inflated to national numbers are more likely to be erroneous when inflating small numbers to national averages. This explains the large discrepancies seen when considering smaller airports in the network. Therefore, OAG can be considered as representing passenger numbers rather than bookings and can be taken for face value, even if validation is strongly encouraged.



(Figure 4.2 continues on next page)

(Figure 4.2 continued)

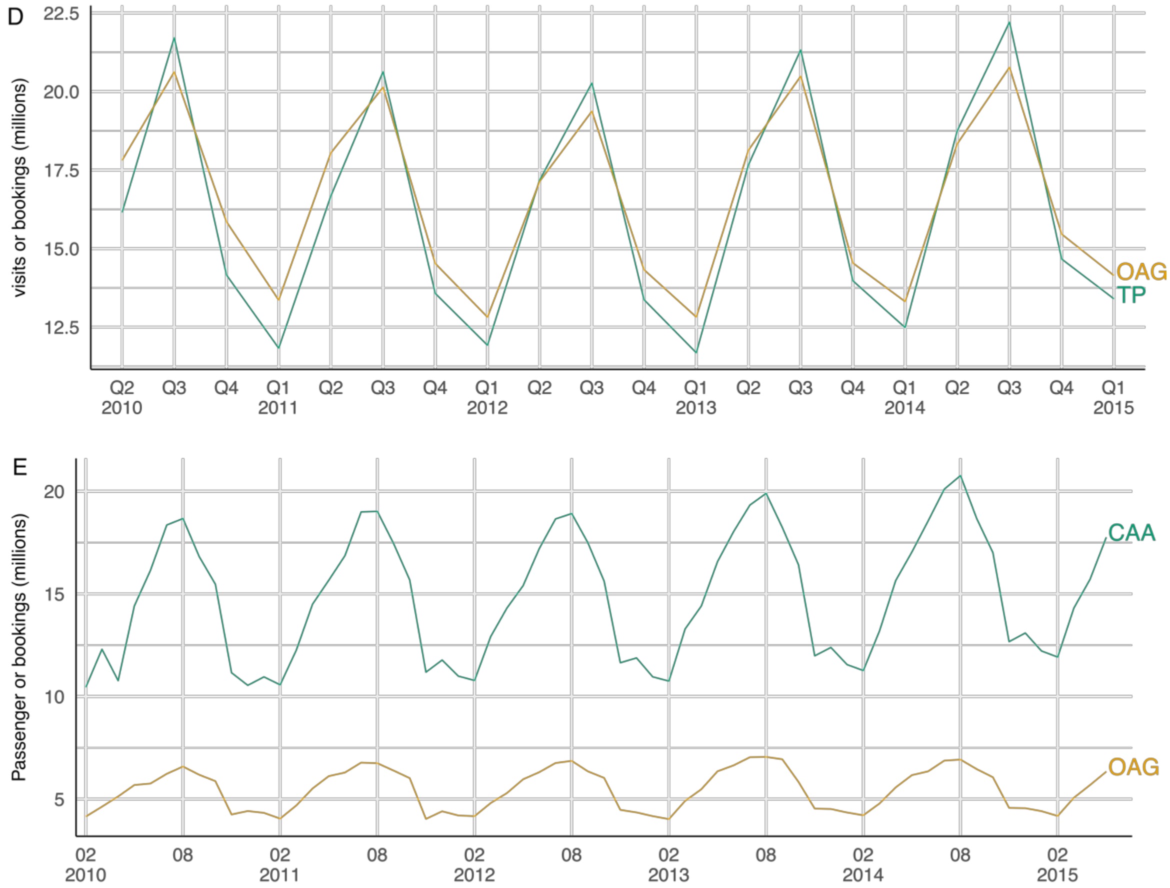
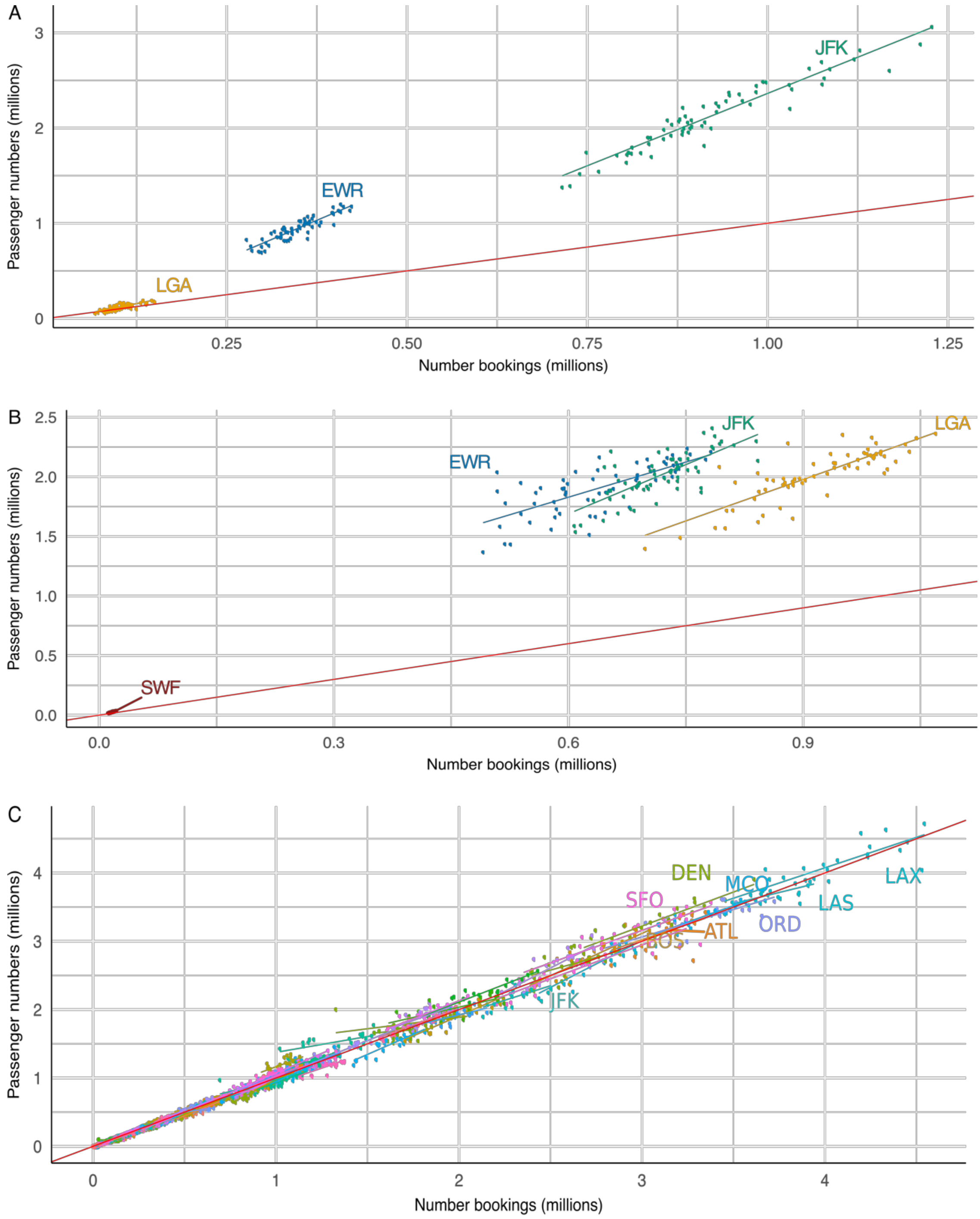
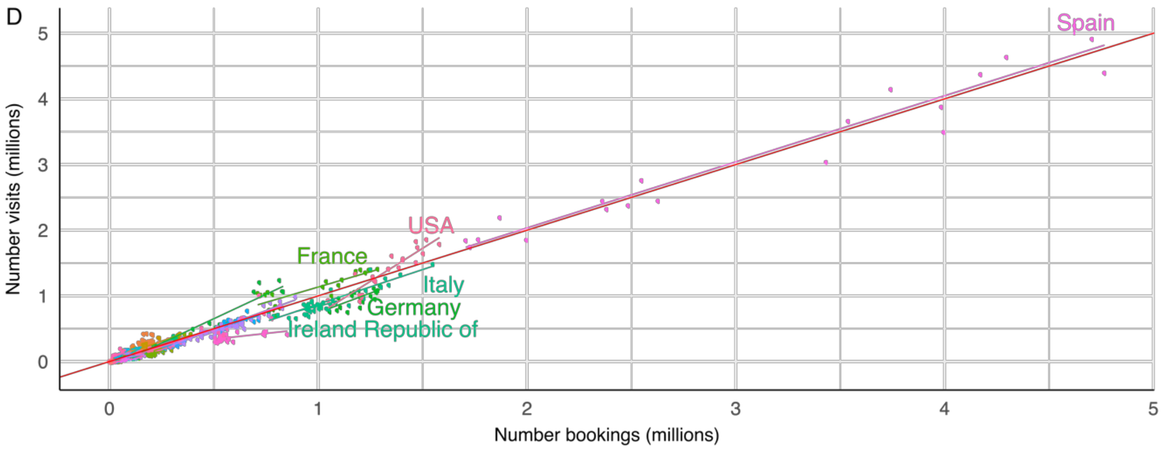


Figure 4.2: Correspondence of temporal patterns between open access data sets, namely (A) PANYNJ international passengers, (B) PANYNJ domestic passengers, (C) USDoT passengers, (D) TravelPac visits, (E) CAA passengers and OAG bookings.



(Figure 4.3 continues on next page)

(Figure 4.3 continued)



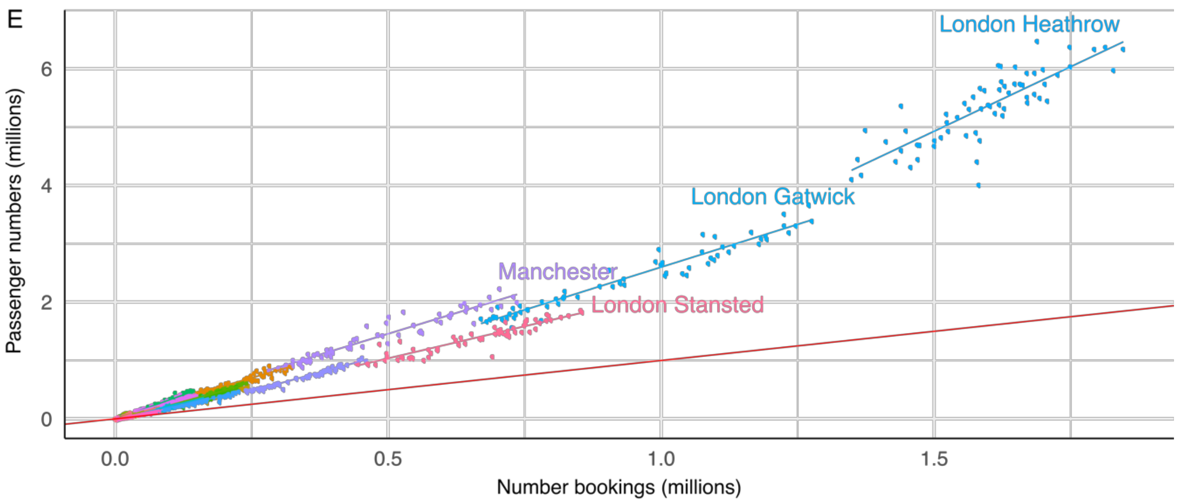
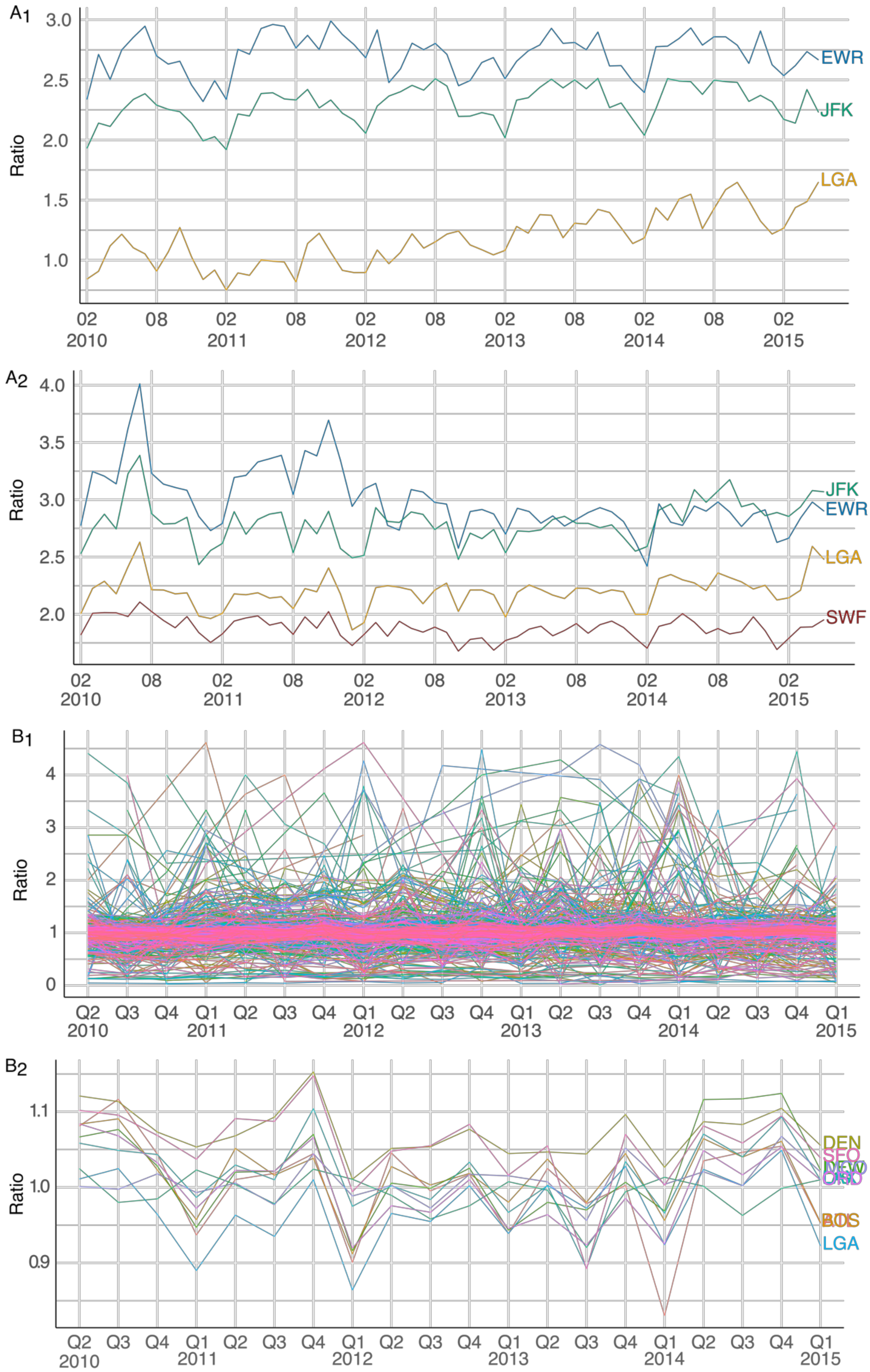


Figure 4.3: Direct comparison between open access data sets, namely (A) PANYNJ international passengers, (B) PANYNJ domestic passengers, (C) USDoT passengers, (D) TravelPac visits, (E) CAA passengers and OAG bookings, by country or airport. Note: the red lines represent the line of equality (x=y) and the coloured lines represent the lines of best fit for each airport or country.



(Figure 4.4 continues on next page)

(Figure 4.4 continued)

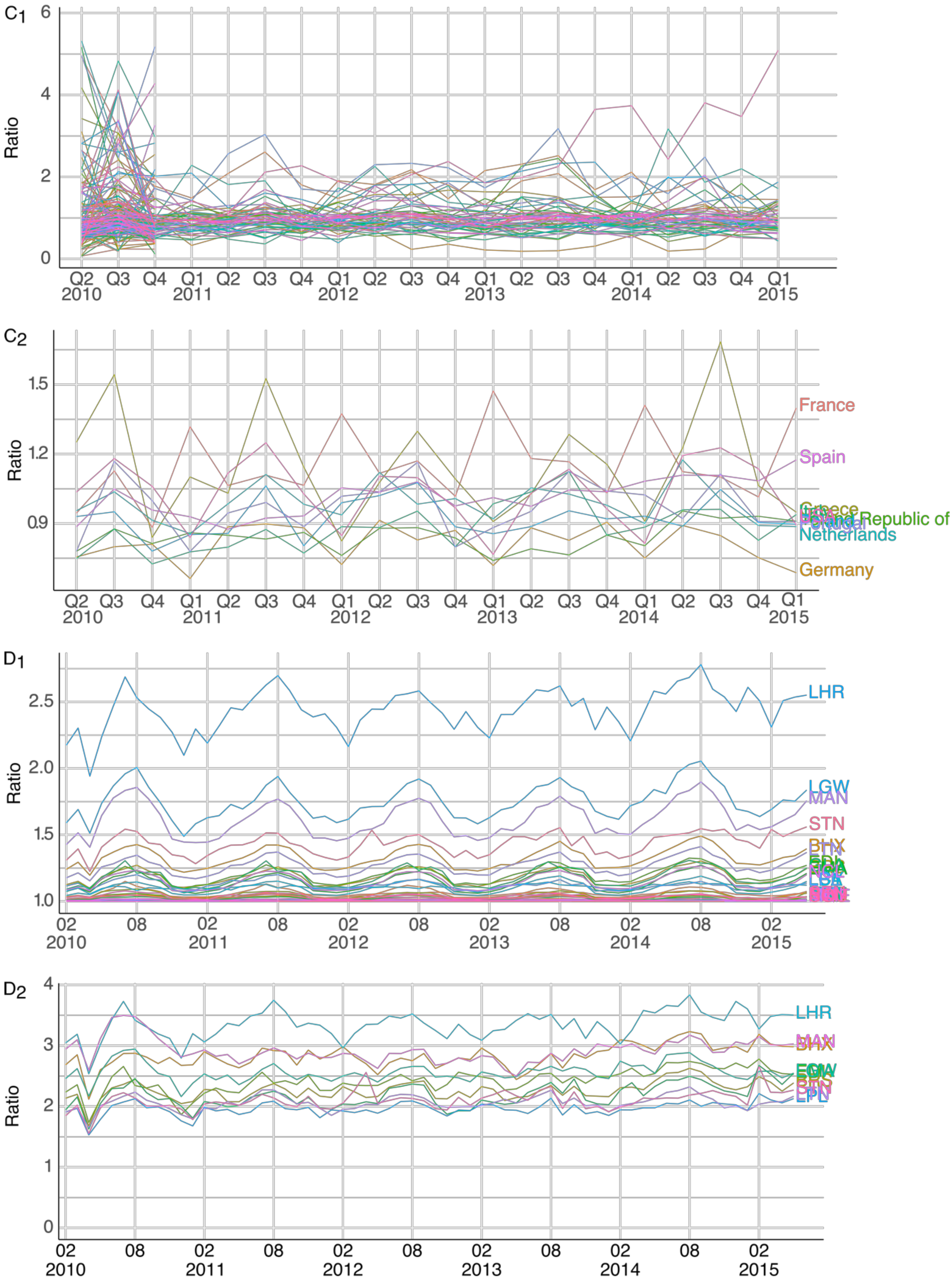
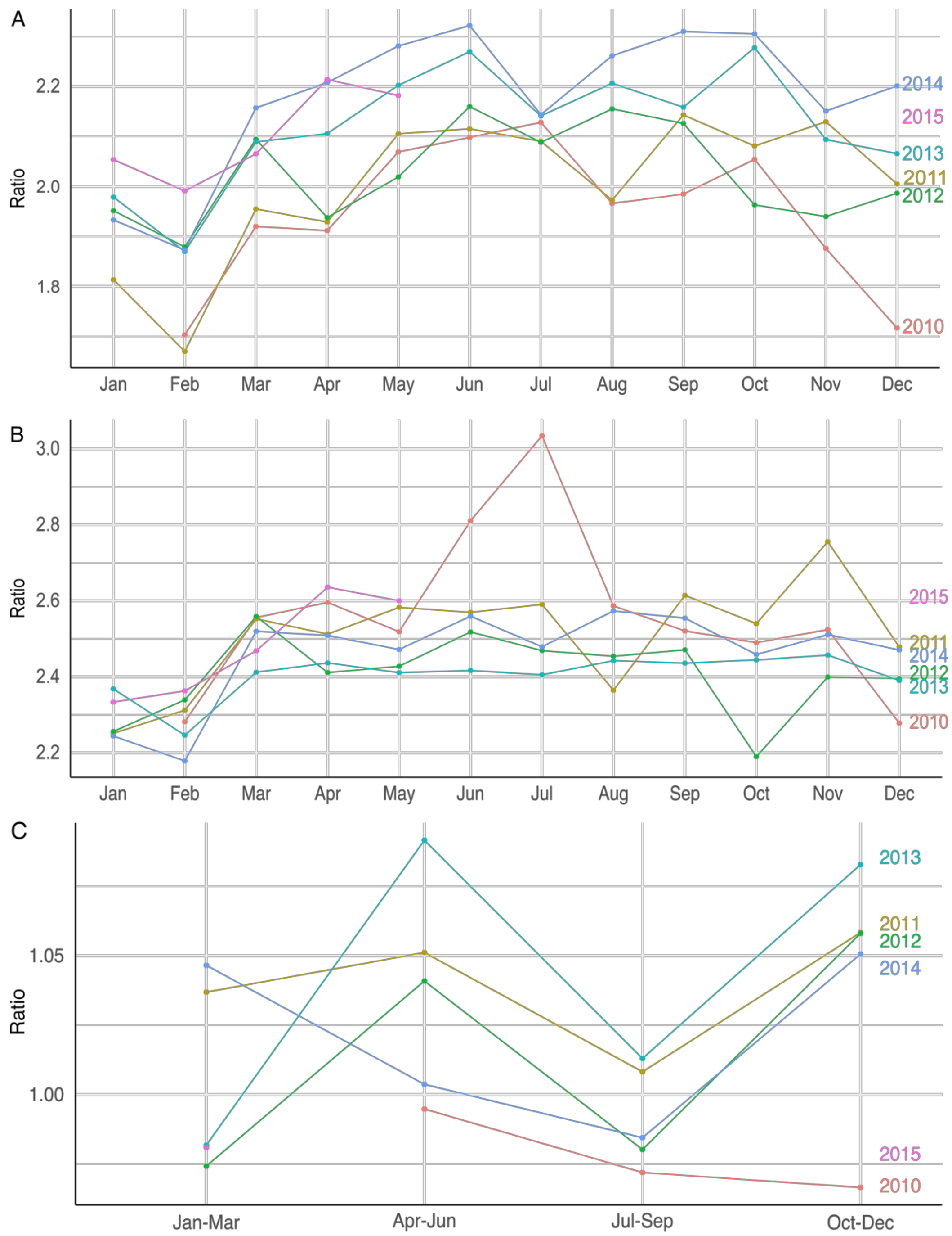


Figure 4.4: Temporal comparison between open access data sets, (A_1) namely PANYNJ international passengers, (A_2) PANYNJ domestic passengers, (B_1) USDoT passengers with ratio values below 5, (B_2) USDoT passengers for selected airports, (C_1) TravelPac visits with ratio values below 7, (C_2) TravelPac visits for selected countries, (D_1) CAA passengers with ratio values below 10, (D_2) CAA passengers leaving selected UK airports and OAG bookings.



(Figure 4.5 continues on next page)

(Figure 4.5 continued)

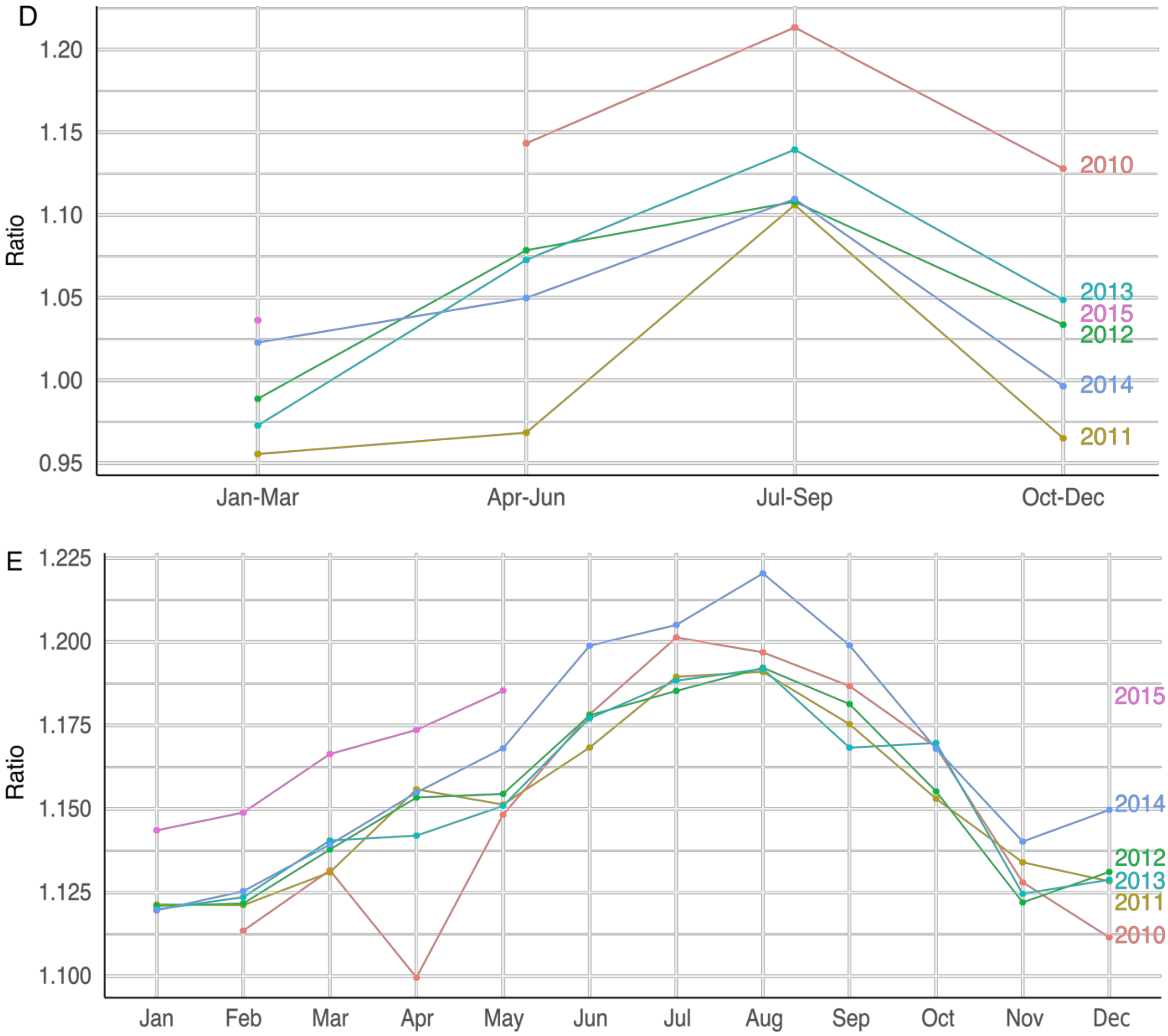
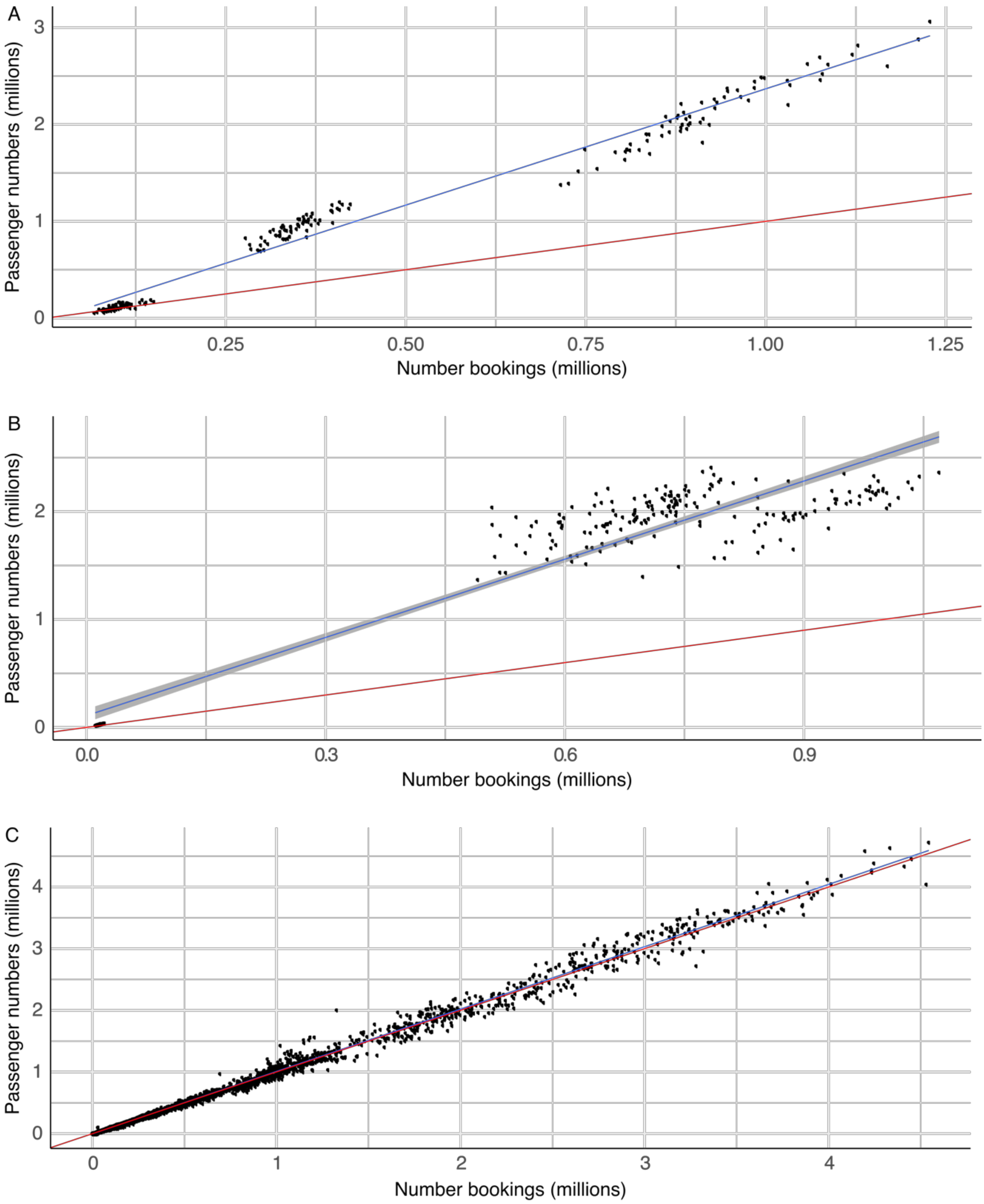
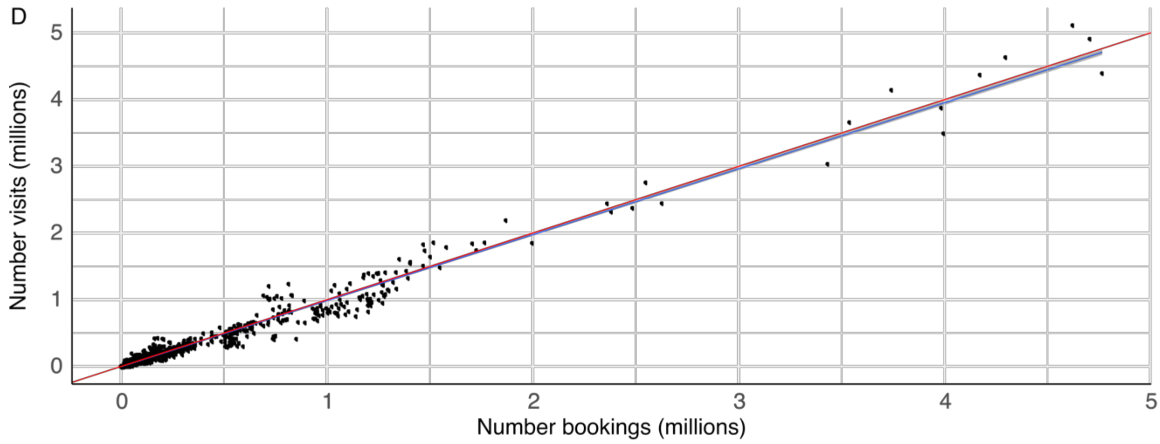


Figure 4.5: Ratios of monthly or quarterly aggregations of open access data sets per OAG bookings by year, namely (A) PANYNJ international passengers, (B) PANYNJ domestic passengers, (C) USDoT passengers (monthly aggregations for ratio values of 5 or less passenger/booking), (D) TravelPac visits, (E) CAA passengers.



(Figure 4.6 continues on next page)

(Figure 4.6 continued)



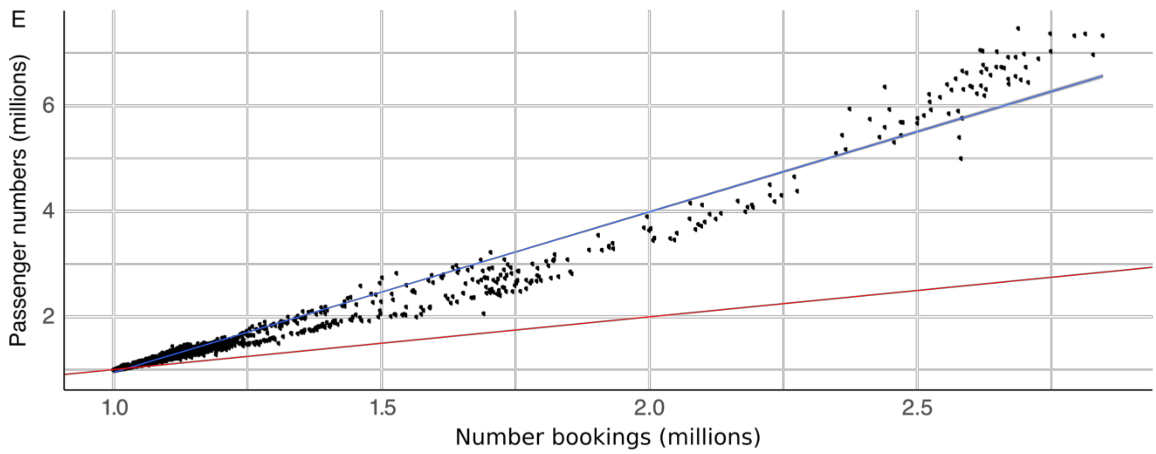
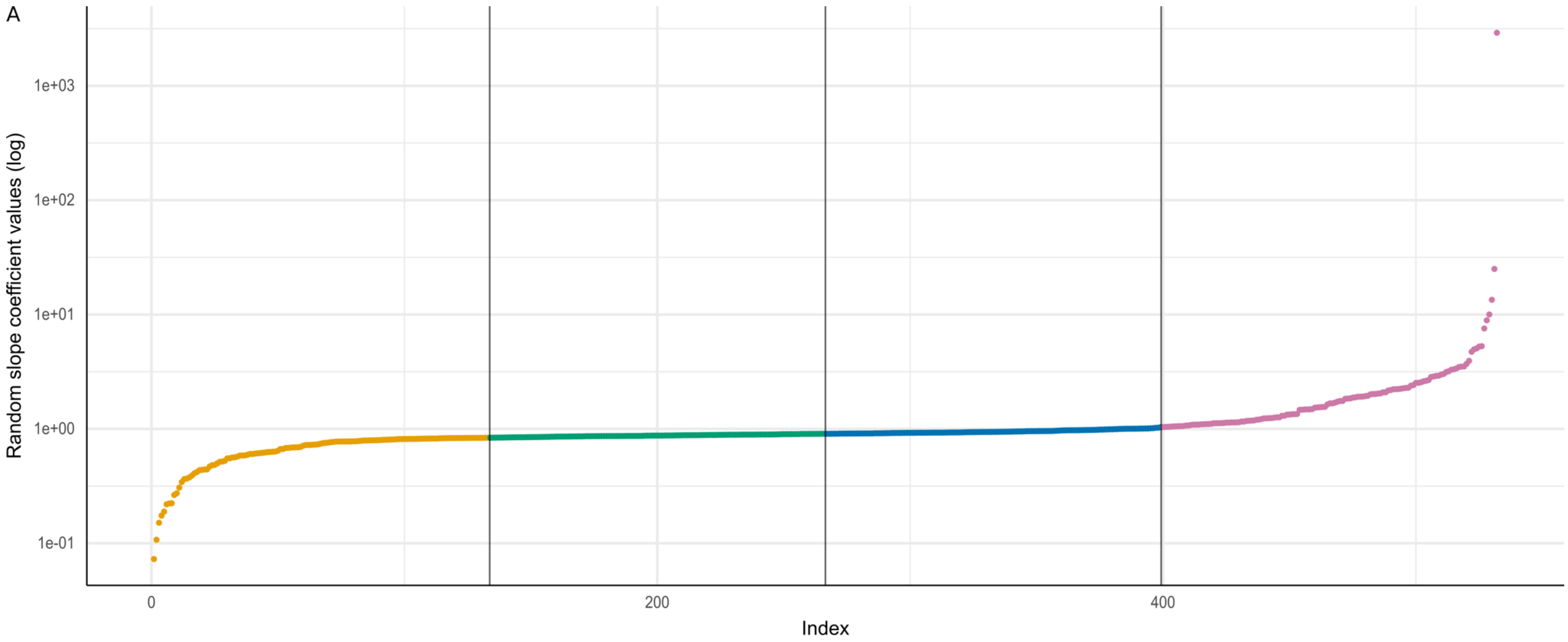


Figure 4.6: Results of the linear regression model using each open access data sets, namely (A) PANYNJ international passengers, (B) PANYNJ domestic passengers, (C) USDoT passengers, (D) TravelPac visits, (E) CAA passengers and OAG bookings.
Note: the red line indicates the line of equality (x=y).



(Figure 4.7 continues on next page)

(Figue 4.7 continued)

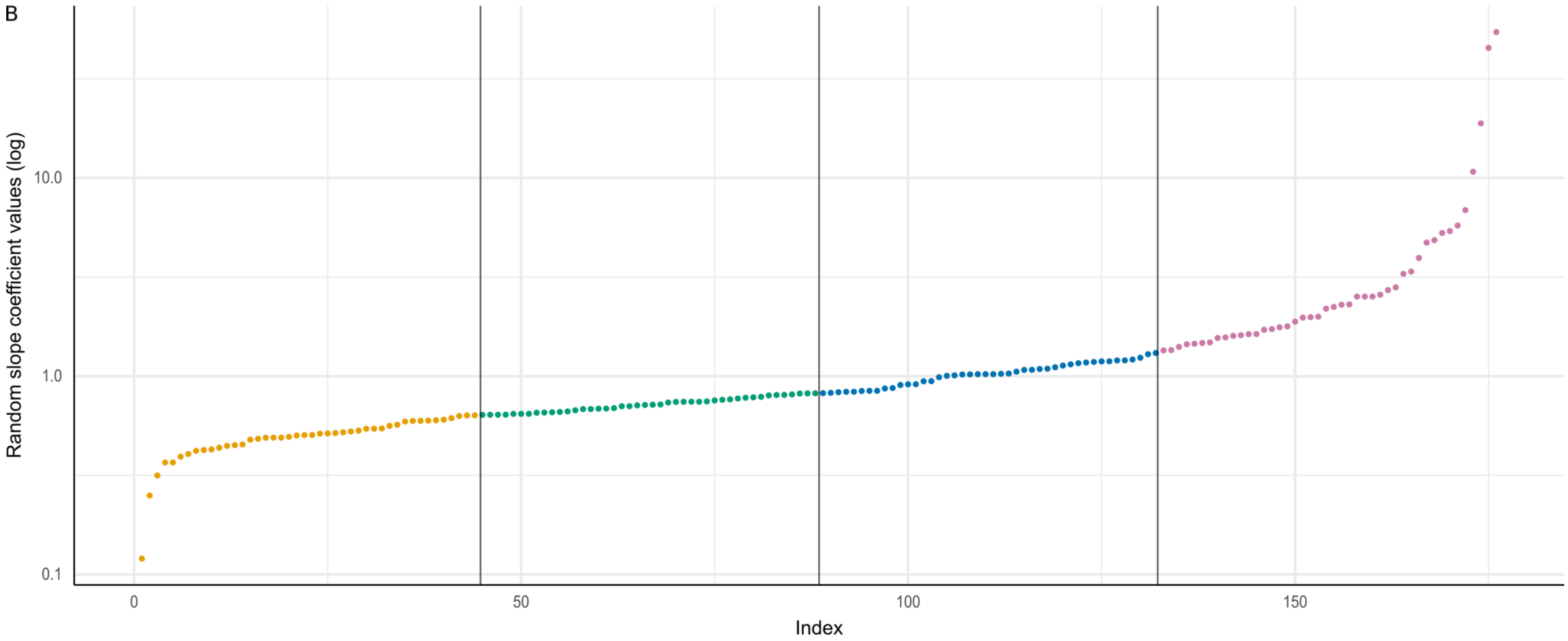
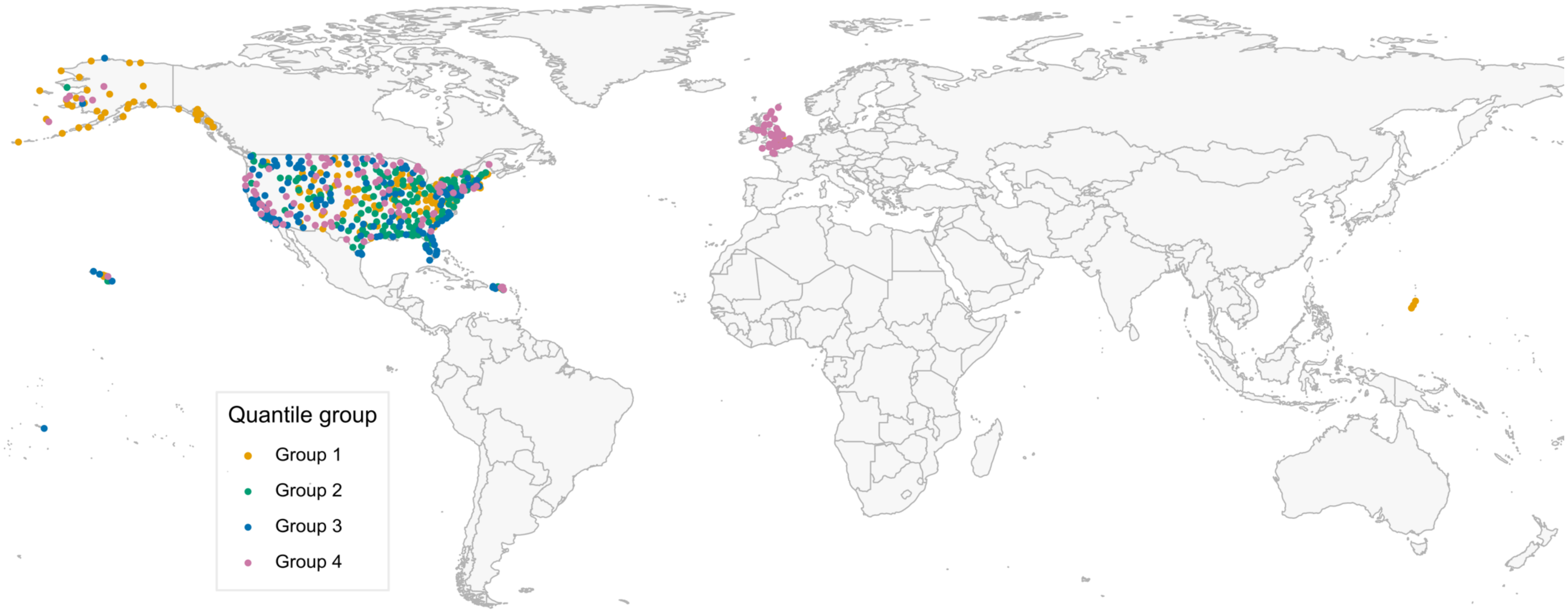


Figure 4.7: Plots representing the value of the slope coefficients (log scale) for each negative binomial model using (A) airports only (data: CAA, PANYNJ, USDoT) and (B) countries (data: TravelPac).
Note: the vertical lines and the different colours of points represent the quartiles of the number of airports and countries included in data, respectively.

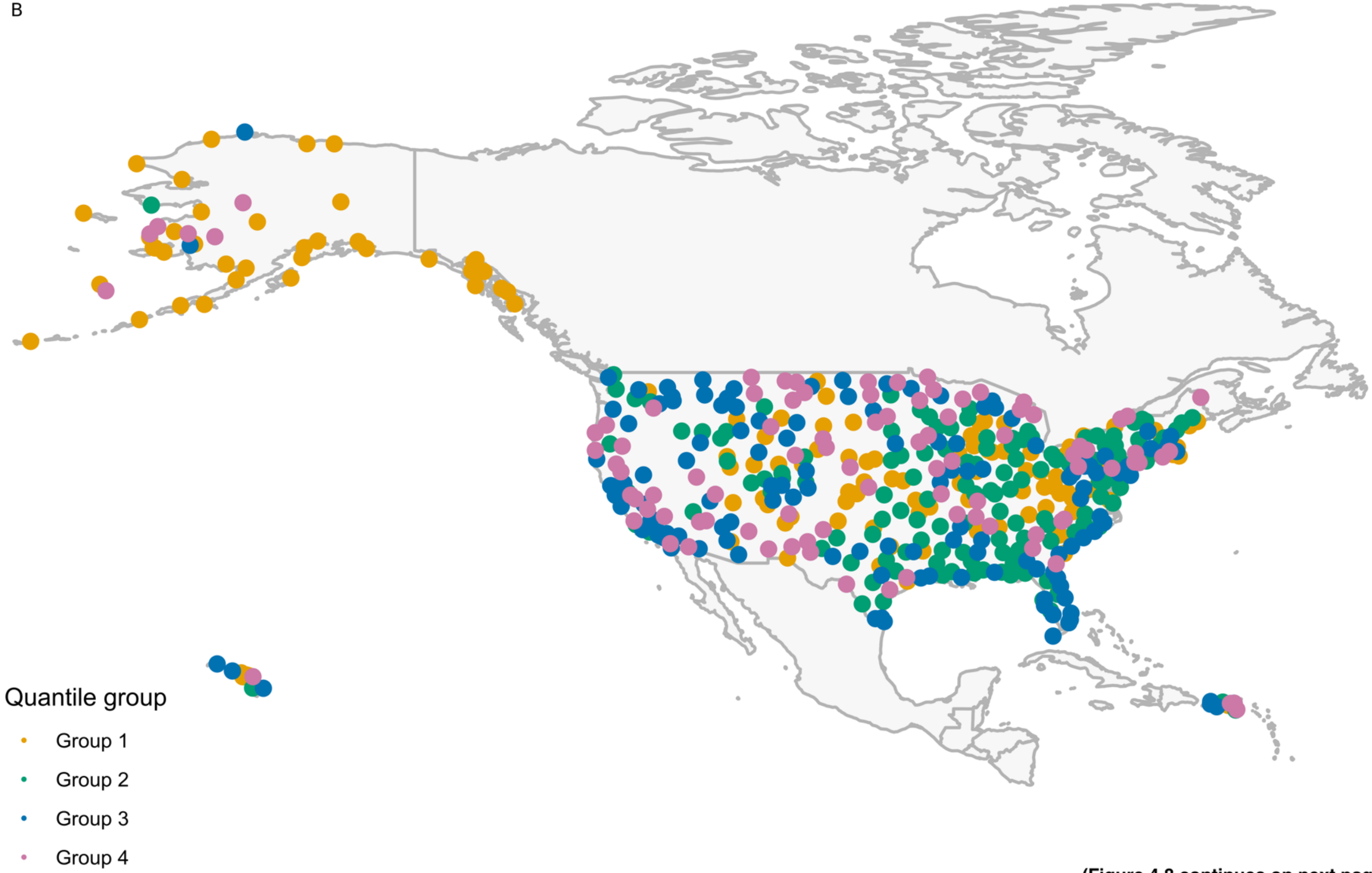
A



(Figure 4.8 continues on next page)

(Figure 4.8 continued)

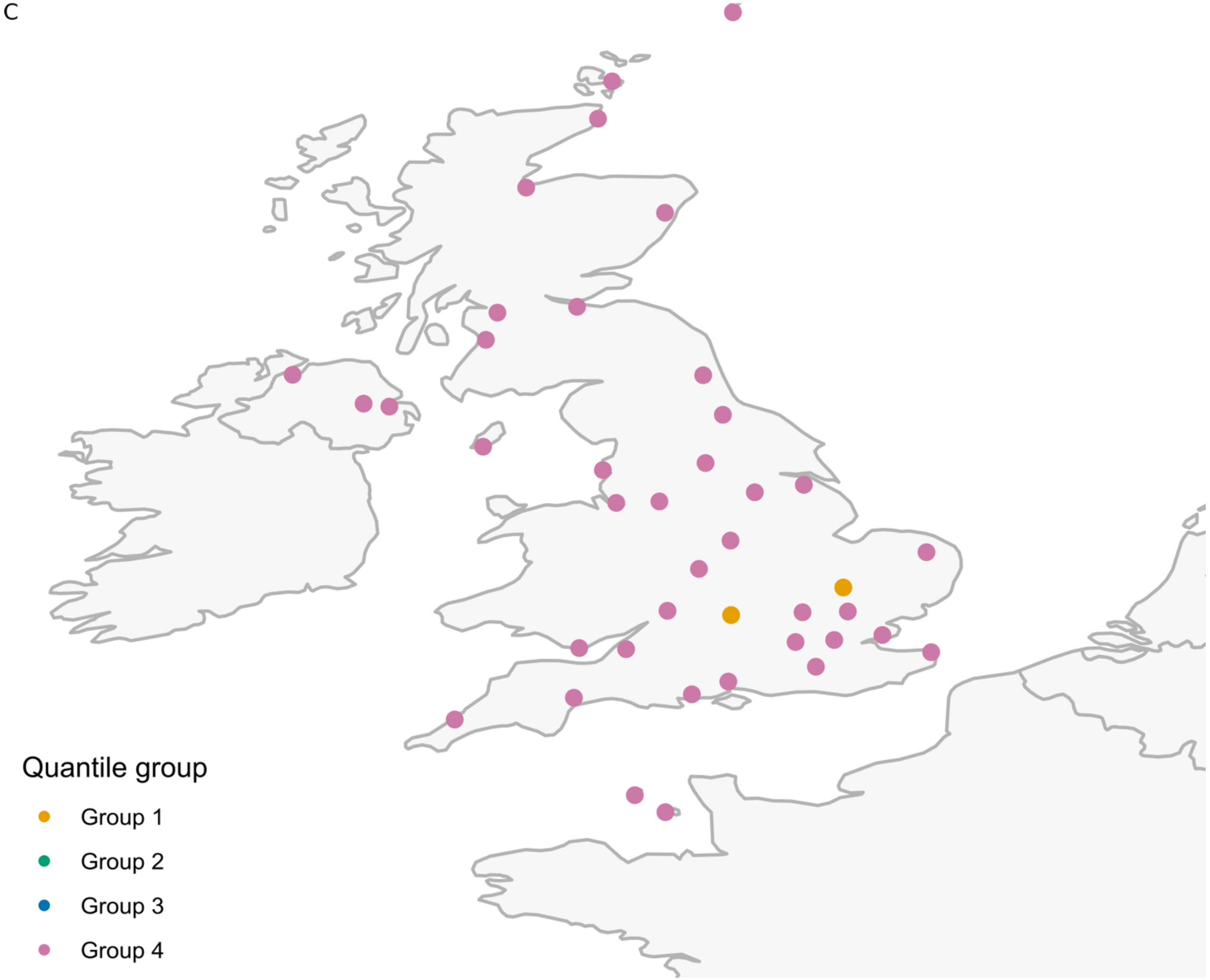
B



(Figure 4.8 continues on next page)

(Figure 4.8 continued)

C



(Figure 4.8 continues on next page)

(Figure 4.8 continued)

D

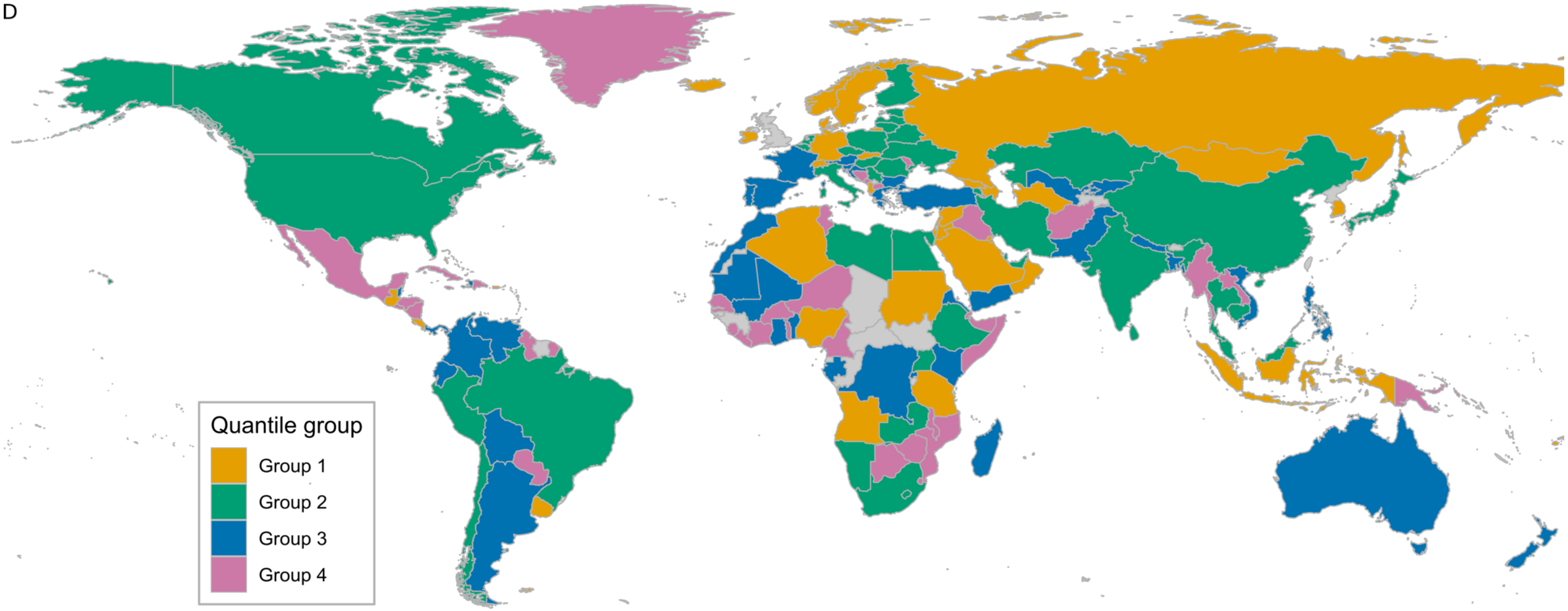


Figure 4.8: Map representations of the location of airports in the (A) world, (B) United States (C) United Kingdom and of countries (D) with random slope coefficients colour coded by quartiles: group 1 representing the first quartile (0-25%), group 2 the second quartile (25-50%), group 3 the third quartile (50-75%) and group 4 the fourth quartile (75-100%).
Note: countries in grey represent the United Kingdom or those without data.

Table 4.2: Summary results of the negative binomial model by airports (total and by data source) and countries, with the number of observations included in each model and the mean of the observations within each quartile, as shown in **Figure 4.7**.

			Slope coefficients							
Geography	Data sets	n	Mean	Median	Min	Max	Quartile 1 (mean)	Quartile 2 (mean)	Quartile 3 (mean)	Quartile 4 (mean)
Countries	TravelPac	1,474	1.88	0.82	0.12	54.36	0.49	0.72	1.03	5.27
Airports	Total	11,415	6.73	0.90	0.07	2,913.30	0.67	0.87	0.95	24.41
	PANYNJ	448	1.04	1.04	0.53	1.36	0.69	1.02	1.23	1.31
	CAA	2,173	1.32	0.93	0.24	10.47	0.61	0.84	1.04	2.85
	USDoT	8,794	7.39	0.97	0.08	3,088.09	0.71	0.94	1.01	26.87

Discussion

The author compared the OAG data against four independent data sets to determine how comparable a commercial data set was with open access data. From this detailed comparison it was possible to determine some similarities and differences between the data sets in the hope of determining which data set matched the OAG subset closest. The author was also able to determine that one OAG booking represented on average about one passenger or visit.

The global airline network had strong seasonal trends with all airports or countries showing most travel occurred on or around August, or Q3, and the least on or around February, or Q1. This is a strong reflection of the holiday periods from the northern hemisphere, where most travel (domestic or international) occurs over the summer holiday period, between July and August. Airline passengers travelling for tourism are motivated by the need to escape from home (Jönsson and Devonish, 2008). Destination weather and holiday availability also played important roles in the trip timing: as shown in the number of trips made to Western Europe during the summer months (June to September) by UK residents, which is double the number of trips made across the rest of the year (Tourism Intelligence International, 2010).

Spain and France as destination countries accounted for the largest number of airline passengers in 2002 (28% of total trips from the UK), and 2004 (Tourism Intelligence International, 2010), and this is reflected in the TravelPac data. As discussed in **Chapter 3**, the USA was the non-EU country from which most international travel to the UK arrived from and accounted for 3.3 million visits in 2015 (Office for National Statistics, 2016; Tourism Intelligence International, 2010).

Similarly to American business travel, business passengers travelling to or from the UK are more likely to travel alone (88% of UK residents travelling for business via Heathrow travel alone). These passengers are also more likely to travel to specific airports such as London City given its proximity to London’s financial district (Civil Aviation Authorities, 2011). Although business travel accounts for a smaller percentage (10%) of purpose of travel than leisure (53%) (US Department of Transport - Bureau of Transportation Statistics, 2015) this purpose of travel may be influencing travel patterns, and transmission of pathogens. It has been shown that passengers travelling for leisure or business will face different infection risks. For example, business travellers have been reported as high risk for sexually transmitted diseases and vaccine preventable diseases (Chen *et al.*, 2018), whereas passengers travelling for tourism were at high risk of diarrhoeal disease (Gautret *et al.*, 2012). Additionally, duration of stay will impact a passenger travelling for leisure's risk of vector borne diseases and diagnosis upon return, with short stays associated with vector borne infection (malaria or dengue for example) or disease, whereas extended travel periods will be linked to tuberculosis infection (Gautret *et al.*, 2012).

When considering using airline passenger data, the importance of knowing what the data contains and its structure cannot be stressed enough. During the comparison between the open access data sets and OAG, some differences became apparent, as previously described. When comparing the inter-monthly variations, for all data sets there was more agreement with OAG during the start and end of the year than during the middle of the year, except for the USDoT where there was more agreement in July to September than April to June and October to December. These graphs showed that during the time periods with most passengers, visits or bookings, there was least agreement between the data sets.

When directly comparing data sets against one another, some large variations could be seen in terms of temporal corresponding data and passengers per booking ratios, with some data sets showing more important variations than others. There was also a large amount of noise within the data themselves. For example, when considering the ratios of passengers per bookings, some airports datasets showed better agreement with OAG (USDoT for example) than others (CAA for example). However, for those data sets with many airports or countries included, a large amount of noise was present. From the negative binomial model, the overall data aggregation by geography was of roughly one airline passenger per booking with important variations present when considering smaller and isolated airports. Data collection methods are likely to play a role in these discrepancies as inflating survey samples to national averages when a small number of passenger travel may lead to over-estimated numbers of passengers travelling. However, passenger purpose of travel is also likely to be impactful, especially for large airports such as London Heathrow and New York John F Kennedy. Given these results, it was not deemed advisable to adjust the OAG data to reflect a preconceived understanding that bookings did not represent true passenger numbers.

In conclusion, when using secondary data to model airline passenger movements, the USDoT and TravelPac are two open access data sets with the best agreement with OAG. Although these are freely accessible, they have a number of drawbacks, including the temporal resolution (quarterly rather than monthly) and geographic (Travelpac being in countries rather than airports; USDoT only considering US domestic travel). The author therefore recommends undertaking one's own comparison analysis against at least one open access data source before undertaking any modelling, to determine the completeness of the data and therefore how appropriate it is for the model considered. However, researchers must be aware of the variations present when directly comparing airports or countries to OAG. Indeed, the passenger numbers linked to some airports and countries have been over or under estimated by OAG, especially in remote locations such as Greenland and Sumburg (UK). However, when undertaking a more appropriate comparison in terms of data size, by combining data from several airports, even from different sources, the overall number of passengers per booking averages to just below one. This can be rounded up to one passenger per booking to have a resulting integer, which is more realistic when considering that the data represents people.

Chapter 5 – Region-specific risk of chikungunya and dengue infection among travellers returning to the United Kingdom

Preamble

Using a subset of the OAG airline data and unique laboratory confirmed anonymised patient data from Public Health England, it was possible to determine which countries posed the highest risk for airline passengers travelling from the United Kingdom. As far as the author is aware, this analysis had not previously been attempted for the United Kingdom regarding chikungunya and dengue specifically nor have studies focused on within-country specific risks for travellers compared to local populations.

Abstract

Assessing the public health risks posed by international travel is an important aspect of planning for novel and emerging infectious diseases (NEID), as airline passengers play a key role in their large-scale spread. Here, the author estimated the relative risk of infection by two different vector-borne viral pathogens (chikungunya and dengue viruses) among travellers returning to the United Kingdom (UK), relative to the residents of the region visited.

Information relating to returning UK travellers diagnosed with dengue or chikungunya was used in combination with contemporary passenger itinerary information. Incidence records for countries visited were gathered to estimate country-specific prevalence. The annual number of UK patients was modelled as a function of within origin-country prevalence and the number of passenger bookings returning from the country to the UK. Region-level effects were fitted to chikungunya and dengue data independently.

After accounting for annual variation in country-specific prevalence, we found several regions where there was a significant difference in the risk of infection relative to the resident population. For both diseases, the relative risk of infection for UK travellers was lowest in the Upper South America compared to the local population. Other regions such as Africa Central (for dengue) and Asia South (for chikungunya) showed point estimates of relative risk suggesting UK visitors were at higher risk than residents.

While effects resulting from systematic reporting biases cannot be excluded, regional-level similarities for two biologically distinct infections suggests our analysis may provide insight into which regions pose varying levels of risk. The author anticipates this information to be useful in parameterising future models of importation risks for vector-borne diseases carried by similar vectors.

Introduction

Every year, vector borne diseases (VBD) cause one billion deaths globally and represent 17% of all infectious diseases cases internationally (World Health Organization, 2017g). Caused by blood-feeding arthropods (including mosquitoes, sand flies and ticks), VBDs affect both rich and poor populations (Centers for Disease Control and Prevention, 2014). Their geographical location and spread has previously been linked to trade, as shown with the international spread of *Aedes albopictus* mosquitoes, a direct result of the global tyre trade (Tatem *et al.*, 2006b). Airline travel has also been linked to the global spread of VBDs by transporting vectors (Tatem *et al.*, 2006b), and infected humans, as was the case in the 2007 chikungunya outbreak in Italy, linked to the return of an infected traveller from India (Angelini R *et al.*, 2007).

Two examples of VBDs are dengue and chikungunya fevers, which will be the main focus of this analysis because of their public health importance and risk of global spread. The dengue virus (family *Flaviviridae*, genus *Flavivirus*) is mainly transmitted by *Aedes (Ae) aegypti* mosquitoes, and also to a lesser extent *Ae albopictus* in the Americas and Europe (World Health Organization, 2014). Although dengue fever is the main cause of childhood hospitalisation in South East Asia, up to 70% of patients do not seek medical treatment, limiting the understanding of the true global burden (Stanaway, 2016). According to World Health Organization (2014), over 40% of the global population (2.5 billion people) lives in areas with a risk of dengue infection, mainly in tropical and urban areas. Dengue infections range from a febrile to haemorrhagic disease with a case fatality rate ranging from 3% to 12% (European Centre for Disease Prevention and Control, 2012). Following an incubation period of four to ten days, symptoms last for up to one week with a high fever, severe headache and nausea. No specific treatment was currently available at time of writing but a vaccine developed by Sanofi Pasteur was registered in early 2016 in several countries (World Health Organization, 2016a).

Historically, chikungunya (family *Togaviridae*, genus *Alphavirus*) caused small outbreaks in rural communities in Africa and Asia. Cases of chikungunya fever have been recorded in Europe, most notably in Italy in 2007 (Rezza *et al.*, 2007) and autochthonous transmissions have been recorded in Southern France (Grandadam *et al.*, 2011). The geographical spread of chikungunya is similar to that of dengue, and the first local transmission in the Americas was reported in Saint Martin in late 2013, starting an outbreak affecting 44 countries (Furuya-

Kanamori *et al.*, 2016). An estimated 1.3 billion people live in at risk areas for chikungunya infection, with countries also reporting co-infections with dengue as both diseases are transmitted by the same vectors (Nsoesie, 2016; European Centre for Disease Prevention and Control, 2012). Following a three to seven day incubation period, symptoms ranging from mild/non-existent to severe illness with high fever and joint pain will appear, usually lasting for days (Public Health England, 2014). Although fatalities are rare, the patient's quality of life can be impaired for months or years with the very young and elderly most at risk of complications (Pan American Health Organization, 2011; Nsoesie, 2016).

Destination and purpose of travel, as well as behaviour while abroad play important roles in a traveller's risk of infection (Vilkman *et al.*, 2016). Tourists sleeping in accommodation with good hygiene and air conditioning, and having visited a travel clinic before travel are at lower risk of catching a VBD while travelling to endemic areas. Conversely, those traveling to visit friends and relatives (VFR) or backpacking are more likely to visit rural environments and therefore at greater risk of VBD (World Health Organization, 2012; World Health Organization, 2001). Travel health professionals often tailor their advice on protection measures based on the travel destination and disease epidemiology; this is in turn informed by World Health Organization (WHO) member states, obliged to alert the global community about potential health threats (Schlangenhauf, 2011).

In May 2015, Zika virus, another VBD carried by *Aedes* mosquitoes, was identified in the Americas, specifically Brazil, for the first time. A causal link between Zika virus and neurological disorders such as microcephaly in new-borns and Guillain-Barré syndrome in adults has since been established (World Health Organization, 2016a). Zika rapidly spread throughout the American continent, but travellers have also transported Zika to geographically distant countries such as Cape Verde (Bogoch *et al.*, 2016).

The aim of this analysis was to determine whether UK travellers were at varying relative risk of VBD infection when travelling abroad compared to local populations in endemic areas, by bringing together patient travel data, airline passenger booking data and passenger duration of travel per country information. This analysis will focus on chikungunya and dengue and may also be applicable to Zika.

Methods

Data on confirmed cases imported into the UK

Information on dengue and chikungunya cases presenting to UK clinics with laboratory-confirmed infection were provided by Public Health England (PHE). Case records were collected from the UK for both diseases. Case specific information included month of laboratory confirmed diagnosis, age group, sex, and recent international travel, including travel destination (countries and/or regions). The data contained information on cases with laboratory confirmed infection between January 2009 and December 2014 inclusive, overlapping the time period for which international airline traveller data was available. The case records were restricted to February 2010 to December 2014, to match the airline data, giving a total of 385 chikungunya and 1,562 dengue cases. Where the destination region name was provided but not the destination country ("Borneo" or "Caribbean", for example), an appropriate region name was assigned. For the purposes of this analysis, 'region' will refer to a group of neighbouring countries, chosen at the discretion of the authors, allowing for model fitting (see below). Surveillance of chikungunya is passively done in the UK, so the travel history available is that reported by the point of care clinician (Public Health England, 2015). After anonymization and cleaning the data, cases with missing or no recent international travel history were excluded from the analysis, as were cases without valid travel destination information for which no region could be assigned. Cases with multiple assigned destination regions were resampled, and randomly assigned to only one of the regions, multiple times, by bootstrapping. Thus, for each bootstrap sample of case data the region in which infection occurred was imputed for all cases. However, a number of cases (n=4 chikungunya and n=59 dengue) had travelled to countries belonging to more than one region and were therefore ignored in the analysis as these observations could not be fitted using the model described below. The model could only be run at the regional and annual level, as the limited number of observations available for each country and year did not allow us to fit a model at the country level and the endemic incidence data was only available at the annual level. This data set did not include duration of travel, so this information was imputed from the Office for National Statistics' TravelPac dataset described in **Chapter 4**. The data were first restricted to only include "air" travellers who were "UK residents" returning from a trip abroad. The duration of travel was categorised as: "Nil stay", "1-3 nights", "4-13 nights", "14-27 nights", "28-90 nights", "3-6 months", "6 months – year" and "Stay not known”, therefore, the number of nights spent in each country and duration of travel were imputed using a uniform distribution.

Traveller data

Monthly airline booking data, from February 2010 to May 2015 inclusive, were extracted from the ‘Traffic Analyser’ database of OAG (OAG, 2016c) previously described. Only itineraries with a destination airport within the UK, an origin airport belonging to countries present in the cases data, and a point of origin airport within the UK were included. Thus, the traveller data represents the number of passenger bookings relating to return travel to the UK from international countries. Monthly bookings were aggregated to annual, country and regional levels before matching the region names from the traveller data to those in the case data regions. Finally, it was assumed that one booking represented one passenger.

Incidence and population data

For each country visited by cases, the annual population size, as well as the number of cases and deaths for both chikungunya and dengue was collected from a range of online open-data sources (**Table 5.2**). Records of annual indigenous incidence of dengue and chikungunya infection were collated from the World Health Organization’s regional office websites. If the data was unavailable from any WHO regional office websites, the author attempted to identify governmental data sources, where possible. A time restriction of data collected was assigned (2010 to 2014) to match the period of the case and airline data. Annual country population sizes between 2010 and 2014 was collated from the World Bank website or if unavailable, from official governmental data sources for that country. For this analysis, countries were grouped by geography into regions, to permit a greater number of observations per regions than would be available at the country level. Countries were grouped by the author according to their close geographic location and climate similarities.

Statistical analysis

Each country may be given a subscript i and be allocated to a region j. We modelled the number of infected passengers (y_{it}) arriving into the UK from country i in year t as a binomial random variable given the total number of airline bookings arriving (F_{it}) from country i in year t, where:

$$y_{ijt} \sim Binomial(p_{ijt}, F_{it}) \quad \textbf{Equation 1}$$

and p_{ijt} is the prevalence of infected individuals among the passengers, noting that this probability is *region* specific. Furthermore, as p_{ijt} is small it was assumed:

$$p_{ijt} = 1 - exp(-\zeta_j * \pi_{it}) \quad \textbf{Equation 2}$$

where π_{it} is the within-country prevalence of infection, and ζ_j is a region-specific parameter, which captures variation in surveillance accuracy, travel behaviour related to exposure risk, and other unobserved between-region heterogeneity. For a specific region we may drop the j subscript such that ζ therefore relates the prevalence of disease in travellers to that of the resident population of the given region: values of 1 indicate that disease prevalence in travellers is the same as that of the resident population. The corresponding likelihood function in this situation is given by:

$$L(\zeta; \pi, F, y) = \prod_{i \in j,t} (\zeta \pi_{it})^{y_{it}} \prod_{i \in j,t} (1 - \zeta \pi_{it})^{F_{it} - y_{it}} \quad \textbf{Equation 3}$$

The approach taken here is knowingly an approximation of a complex system, and strikes a balance between data availability and model parsimony. Maximum likelihood estimates for ζ were obtained for each region using the Brent optimisation within R's optim() routine (R Core Team, 2017), with asymptotic confidence intervals derived from the Hessian matrix.

A second model was run to include the duration of travel for passengers to each country, imputed from TravelPac data. Only the year 2010 only was used, as this had the largest number of observations and we assumed that duration of travel stayed constant throughout each year and between years. As the incidence of both diseases was small, it was assumed:

$$p'_{ijtd} = 1 - exp\left(-\zeta_j * \pi_{it} * (\frac{d_j}{365})\right) \quad \textbf{Equation 4}$$

p'_{ijtd} was used as described previously to run the same model even when duration of travel was included, with d_j the duration of travel in days imputed from TravelPac data, π_{it} is the within-country prevalence of infection and ζ_j is a region-specific parameter. We assumed that the duration of travel per country did not change depending on the time of year as seasonality was not included in the model. Similarly, the likelihood function for the model including duration of travel was:

$$L'(\zeta;\pi,F,y,d) = \prod_{i\in j,t}(\zeta\pi_{itd})^{y_{it}}\prod_{i\in j,t}(1-\zeta\pi_{itd})^{F_{it}-y_{it}} \qquad \textbf{Equation 5}$$

With the Maximum likelihood (ζ) for each region were estimated using the same optimisation method described for Equation 3.

Finally, the absolute risk per country was also calculated by dividing the number of imported cases from each country by the number of corresponding returning passengers.

Results

Data description

Countries and regions visited

From the 432 chikungunya cases recorded in the United Kingdom between February 2010 and December 2014, 19 (4%) had no associated travel history and were excluded from the analysis. A further 30 (7%) had no destination, clear country or region stated, and were also excluded. An additional two cases were excluded as they had travelled to two distinct regions and did not fit into the regional model, therefore a total of 382 cases remained and were included in the model. Of the 1,941 dengue cases reported between February 2010 and December 2014, only one dengue case reported not having travelled, and 367 (19%) cases were recorded as "not stated" in the travel field, and therefore excluded. From the 1,573 cases remaining, 19 reported a region name from which no country or region could be assigned ('Latin America' for example which is not politically recognised) and were also excluded. A total of 49 (3%) cases had travelled to countries belonging to more than one region, and did not fit the regional model so were also excluded, leaving 1,505 dengue cases included in the model.

The total number of cases that travelled to each region was highlighted in **Figure 5.1**. Jamaica was the most visited country by chikungunya cases, with all visits (n=90, 24%) occurring in 2014, and India (n=89, 23%) was the second most visited country by chikungunya cases, across all years available. For dengue, India and Thailand were the most visited countries with n=362 (24%) and n=361 (24%) cases, respectively. The majority of chikungunya cases had visited Caribbean countries (n=219, 57%). However, the majority of dengue cases (n=784, 52%) travelled to South East Asia (especially Thailand), followed by South Asia (especially

India) (n=578, 48%). The UK saw a total of 23 dengue imported cases from Western Europe (specifically Madeira, Portugal), and only in 2012.

There was an overall increasing and seasonal trend in the number of both chikungunya and dengue cases seen, with dengue cases peaking on or around August and chikungunya cases peaking around November each year (**Figure 5.2 A**). Although the number of chikungunya cases remained lower than those of dengue before 2014 (up to 19 chikungunya cases per month), a sharp rise can be seen from June 2014 onward (n=275 in 2014 alone compared to n=160 between 2010 and 2013), which was not seen in previous years. The total number of imported cases by month showed no strong correlation with the airline passenger numbers according to country and time series (**Figure 5.2 A**).

The variation of age distribution of cases differed according to disease (**Figure 5.2 B**)**,** with the majority of dengue cases recorded in the 20-24 age group (n=238, 15%) and the majority of chikungunya cases in the 35-39 age group (n=55, 14%). However, when breaking down each age group by sex, the majority of cases diagnosed with chikungunya virus were female (n=245, 58%) and between 55 and 59 years old (n=35). Fewer males were diagnosed with chikungunya (n=175, 42%). With regards to dengue cases, the majority were male (n=860, 55%) and aged between 30 and 34 years old (n=113). Of the female cases (n=686) the 20 to 24 age group was most represented (n=110).

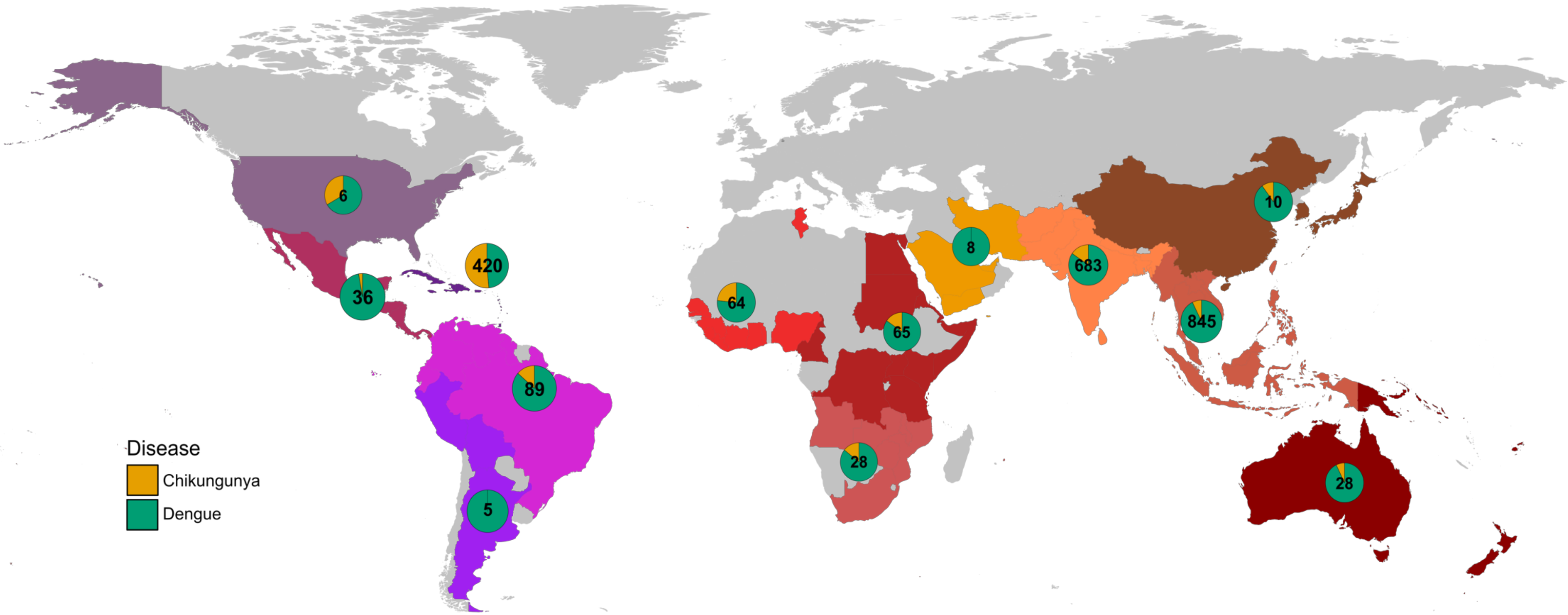
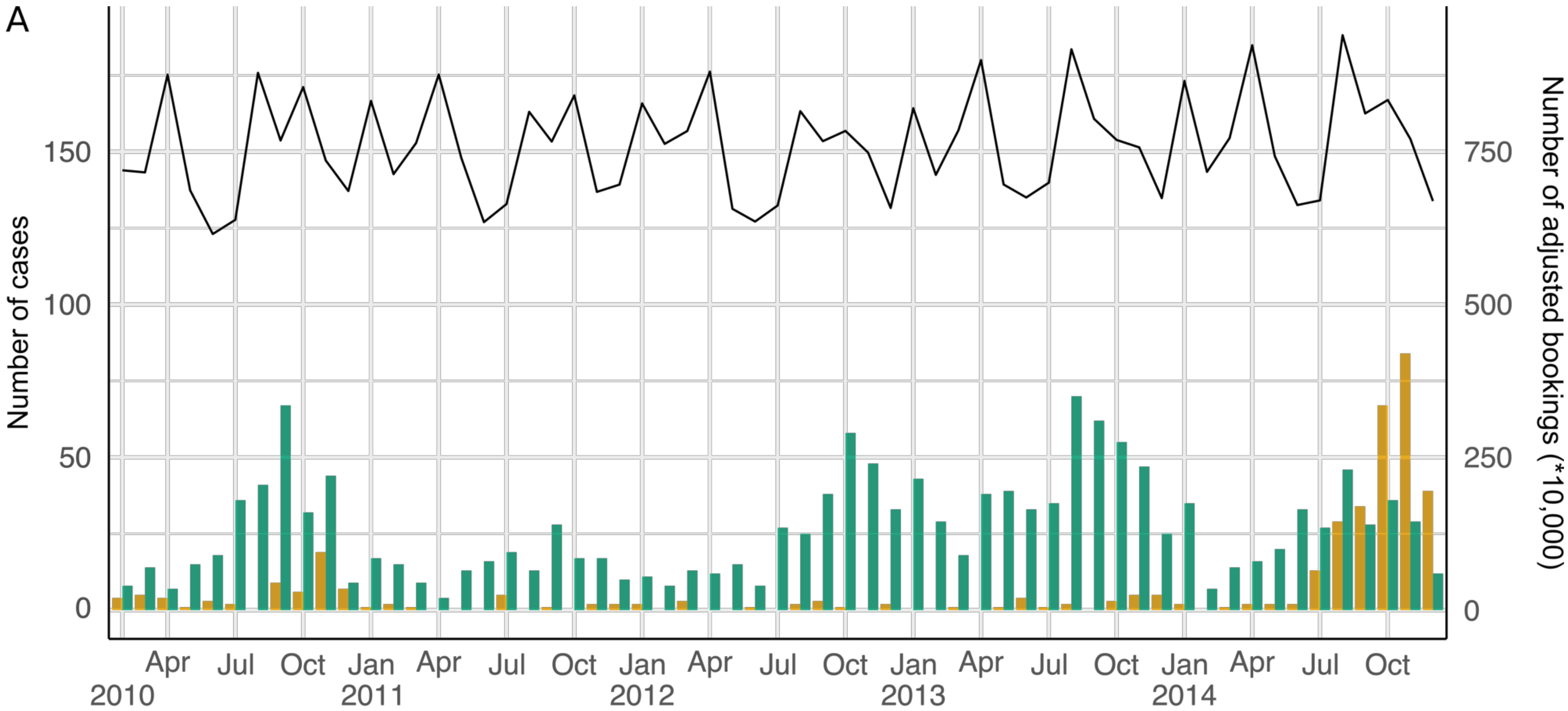


Figure 5.1: Map of countries visited by cases diagnosed with chikungunya and dengue after international travel between 2010 and 2014 grouped by region (see Table 5.2). Note: the Portuguese island of Madeira was included in the West Africa region, due to geographical proximity.



(Figure 5.2 continues on next page)

(Figure 5.2 continued)

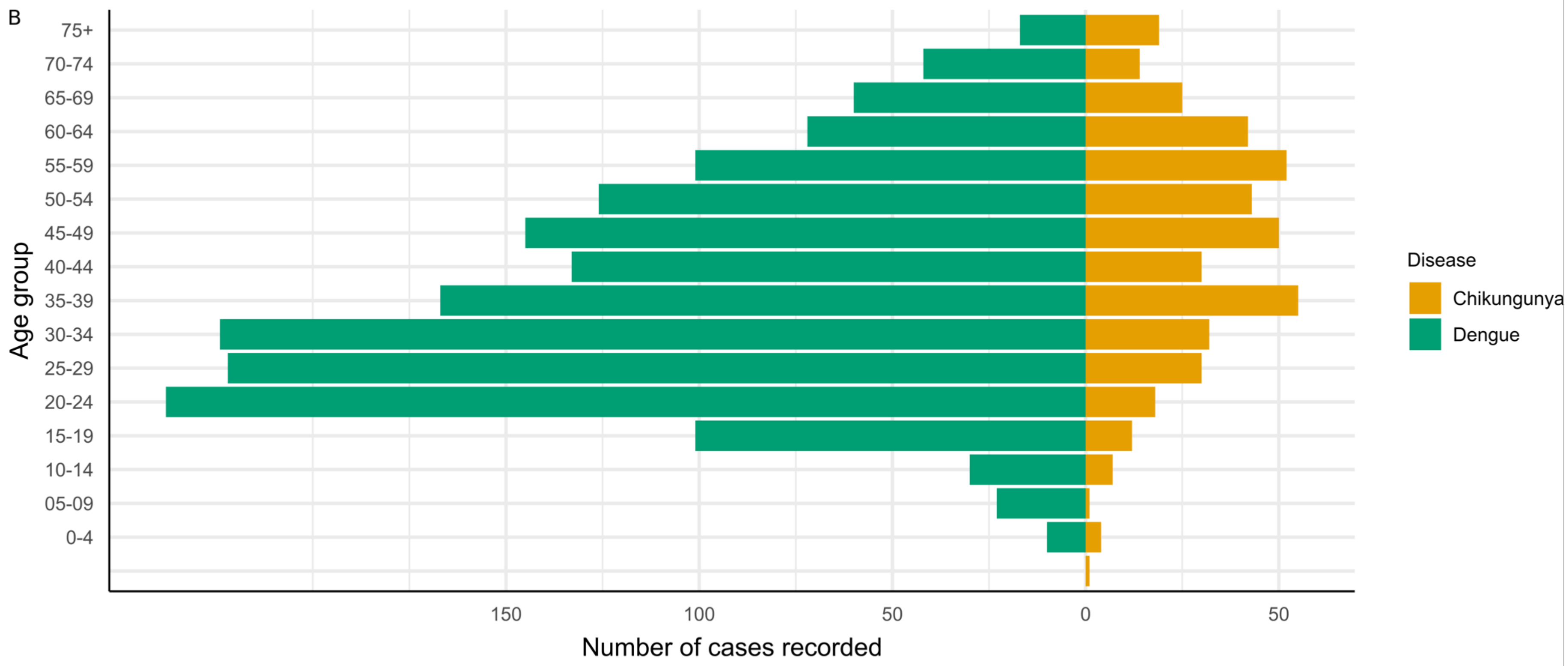


Figure 5.2: (A) Time distribution of international travel-associated chikungunya and dengue cases reported in the United Kingdom, with the black line showing the number of bookings returning from visited countries; (B) age distribution of chikungunya and dengue cases reported in the United Kingdom.

Results from model fitting

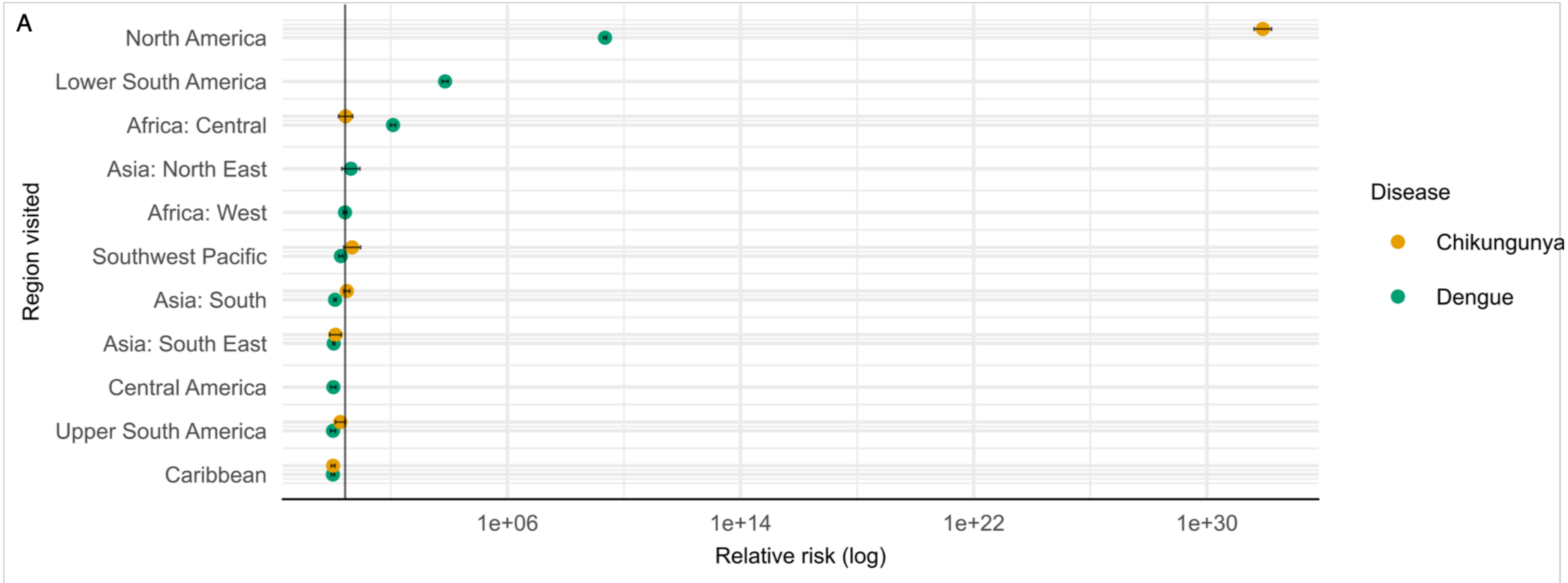
The model was run at an annual and regional level for each disease to determine the relative risk of travellers becoming infected within each region, compared to the local populations. By aggregating countries to the regional level, the model worked more effectively, by overcoming the problem of some countries having very few observations, and therefore the model not fitting (data not shown).

When not considering duration of travel, travellers visiting the Caribbean were at a reduced relative risk of becoming infected with either disease compared to the local populations (**Figure 5.3 A**; coefficients in **Table 5.1**). Additionally, regions such as Upper South America, Central America and South East Asia also represented a reduced relative risk for UK travellers of becoming infected with dengue compared to the local populations (coefficients in **Table 5.1**). Travellers to North America were at a significantly increased relative risk of chikungunya infection compared to the local population. Lastly, regions such as Central Africa, North America and Lower South America also posed a higher relative risk for UK visitors of becoming infected with dengue.

In contrast, when duration of travel was included in the model, some variations were seen relative to the level of risk encountered by travellers compared to local populations (**Figure 5.3B**, **Table 5.1**). When comparing the overall risks for each disease, the average risk of infection with chikungunya was estimated at $4.09e^{+03}$ and $1.18e^{+31}$ with and without considering duration of travel, respectively. When considering the risk of infection with dengue, the average risks were 21.86 and $2.034e^{+08}$ with and without considering duration of travel. These averages showed a protective effect when including duration of travel in the model. When considering global regions, passengers travelling to the same regions as earlier (Caribbean, Upper South America and Central America) also faced a reduced risk compared to local populations. On the other hand, for regions such as Southwest Pacific, Asia South and Asia South East, travellers were at increased risk of infection with chikungunya, whereas Africa central and Asia South were riskiest regarding dengue infection, compared to local populations. From this comparison, it was clear that including purpose of travel to the model had an impact on the level of risk faced by UK travellers compared to local populations.

Overall, the absolute risk for UK travellers was low (**Figures 5.4** and **5.5**), with visited country mean of $5.16e^{-05}$ (confidence intervals ranging between $4.92e^{-05}$ and $5.41e^{-05}$) and region mean of $5.17e^{-05}$ (confidence intervals of $4.62e^{-05}$ and $1.02e^{-04}$). Tonga (South West Pacific) posed the highest risk for chikungunya (0.0081, 95% CI 0.00021 and 0.044) whereas

Guadeloupe (Caribbean) posed the highest absolute risk for dengue (0.0045, 95% CI 0.0012 and 0.011). It must be noted that the total number of airline bookings for each country over the time period considered was small (123 for Tonga and 891 for Guadeloupe) (Guadeloupe is a French region but was considered as a country for the purposes of this analysis). Variations were also seen between regions according to disease: the Caribbean posed the highest absolute risk for chikungunya but not for dengue ($4.49e^{-04}$ for chikungunya with 95% CI $3.90e^{-04}$ and $5.13e^{-04}$; $8.68e^{-05}$ for dengue and 95% CI $7.50e^{-05}$ and $1.00e^{-04}$). On the other hand, South East Asia saw the highest absolute risk for dengue but not for chikungunya: $1.13e^{-03}$ for dengue (95% CI $1.04e^{-04}$ and $1.24e^{-04}$) and $2.05e^{-05}$ (95% CI $1.38e^{-05}$ and $2.95e^{-05}$).



(Figure 5.3 continues on next page)

(Figure 5.3 continued)

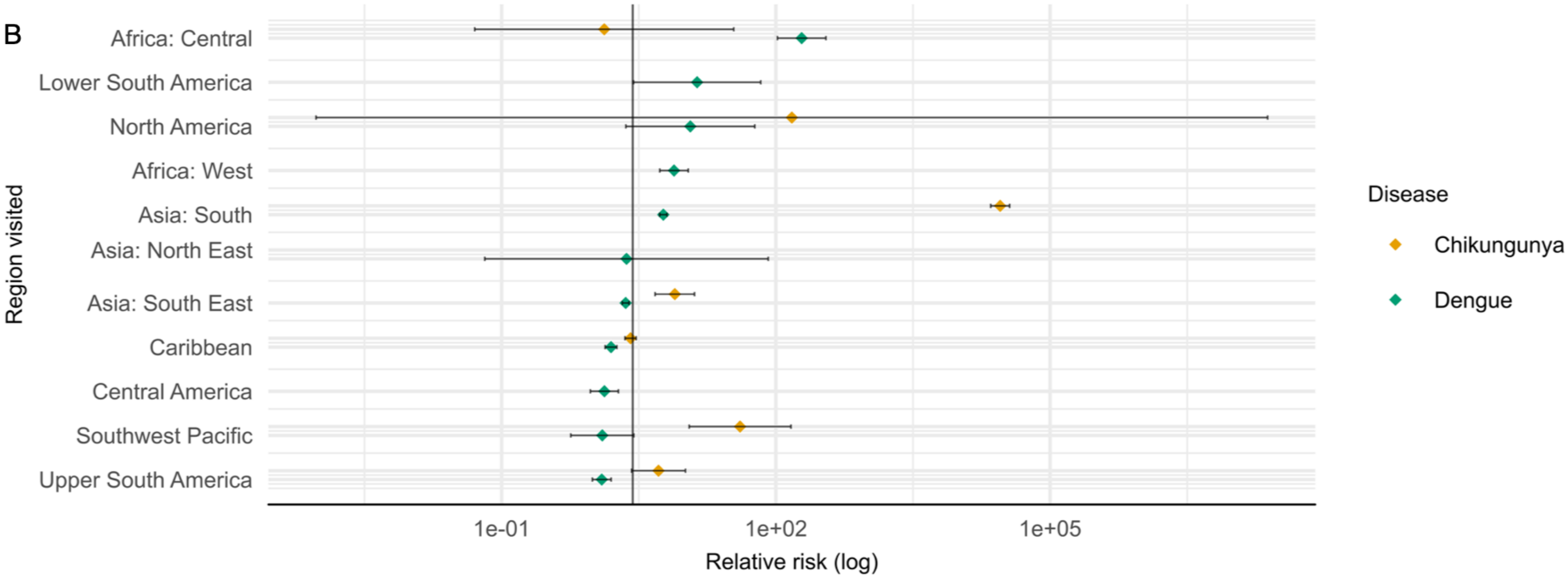


Figure 5.3: (A) Not considering duration of travel, regional relative risk (of infection prevalence for airline passengers relative to within-country population) model results and (B) considering duration of travel, regional relative risk (of infection prevalence for airline passengers relative to within-country population) model results; both for returning cases of chikungunya and dengue virus cases, with 95% confidence intervals surrounding each point.
Note: some values were so large (see **Table 5.1**) so the log scale was used for clarity. The vertical line represents the line of equality (travellers are at equal risk to the local population).

Table 5.1: Table of coefficient results (ζ in equation 2), standard error, absolute risk values and confidence intervals (95%) for each region, disease and whether duration of travel was included. The absolute risk was calculated as the number of imported cases divided by the number of returning travellers, for the given region.
Note: four regions (Africa South (Chikungunya and Dengue) and Middle East (Dengue)) were included in the absolute risk calculations but in the relative risk, as their endemic prevalence could not be determined.

		No duration included			Duration included				
Region name	**Virus name**	**Relative risk**	**SE**	**95% CI**	**Relative risk**	**SE**	**95% CI**	**Absolute risk**	**95% CI**
Africa: Central	Chikungunya	2.83	1.33	1.62 – 4.92	1.33	5.28	$5.08e^{-02}$ – $3.46e^{+01}$	$4.43e^{-05}$	$1.63e^{-05}$ - $9.64e^{-05}$
Asia: South	Chikungunya	3.10	1.12	2.47 – 3.89	$2.84e^{+04}$	1.13	$2.24e^{+04}$ – $3.60e^{+04}$	$1.99e^{-05}$	$1.58e^{-05}$ – $2.49e^{-05}$
Asia: South East	Chikungunya	1.27	1.26	$8.02e^{-01}$ – 2.00	7.85	1.29	4.77-12.90	$2.05e^{-05}$	$1.38e^{-05}$ – $2.95e^{-05}$
Caribbean	Chikungunya	1.05	1.07	$9.14e^{-01}$ – 1.21	2.57	1.07	2.24 – 2.96	$4.49e^{-04}$	$3.90e^{-04}$ – $5.13e^{-04}$
North America	Chikungunya	$8.28e^{+31}$	1.42	$4.17e^{+31}$ – $1.64e^{+32}$	$1.50e^{+02}$	452.60	$9.32e^{-04}$ – $2.40e^{+07}$	$3.10e^{-07}$	$7.86e^{-09}$ – $1.73e^{-06}$
Southwest Pacific	Chikungunya	4.72	1.41	2.42 – 9.20	40.56	1.92	$1.13e^{+01}$ – $1.46e^{+02}$	$5.66e^{-06}$	$6.85e^{-07}$ – $2.04e^{-05}$
Upper South America	Chikungunya	1.86	1.22	1.26 – 2.73	5.20	1.41	2.64 – $1.02e^{+01}$	$3.13e^{-04}$	$1.50e^{-04}$ – $5.76e^{-04}$
Africa: Central	Dengue	$1.20e^{+02}$	1.11	$9.71e^{+01}$ – $1.47e^{+02}$	$1.92e^{+02}$	1.37	$1.04e^{+02}$ – $3.53e^{+02}$	$2.51e^{-05}$	$1.79e^{-05}$ – $3.44e^{-05}$
Africa: Western	Dengue	2.69	1.07	2.36 – 3.08	7.69	1.20	5.37 – $1.10e^{+01}$	$2.09e^{-05}$	$1.49e^{-05}$ – $2.86e^{-05}$
Asia: North East	Dengue	4.26	1.44	2.10 – 8.66	2.32	6.18	$6.54e^{-02}$ – $8.25e^{+01}$	$4.54e^{-06}$	$1.15e^{-07}$ – $2.53e^{-05}$
Asia: South	Dengue	1.23	1.05	1.12 – 1.34	5.87	1.05	5.36 – 6.44	$7.75e^{-05}$	$7.07e^{-05}$ – $8.47e^{-05}$

(Table 5.1 continues on next page)

(Table 5.1 continued)

		No duration included			Duration included				
Region name	**Virus name**	**Relative risk**	**SE**	**95% CI**	**Relative risk**	**SE**	**95% CI**	**Absolute risk**	**95% CI**
Asia: South East	Dengue	1.10	1.04	1.02 - 1.19	2.27	1.05	2.09 - 2.48	$1.13e^{-04}$	$1.04e^{-04}$ - $1.24e^{-04}$
Caribbean	Dengue	1.03	1.07	$9.11e^{-01}$ - 1.17	1.57	1.08	1.36 - 1.82	$8.68e^{-05}$	$7.50e^{-05}$ - $1.00e^{-04}$
Central America	Dengue	1.08	1.10	$8.85e^{-01}$ - 1.31	1.33	1.20	$9.35e^{-01}$ - 1.89	$4.32e^{-05}$	$2.85e^{-05}$ - $6.29e^{-05}$
Lower South America	Dengue	$7.31e^{+03}$	1.13	$5.77e^{+03}$ - $9.27e^{+03}$	13.75	2.27	2.77 - $6.83e^{+01}$	$4.05e^{-05}$	$1.03e^{-06}$ - $2.26e^{-04}$
North America	Dengue	$2.24e^{+09}$	1.06	$2.00e^{+09}$ - $2.51e^{+09}$	11.59	2.29	2.29 - $5.88e^{+01}$	$2.98e^{-07}$	$7.54e^{-09}$ - $1.66e^{-06}$
Southwest Pacific	Dengue	1.94	1.08	1.66 - 2.28	1.26	1.50	$5.71e^{-01}$ - 2.79	$1.10e^{-05}$	$4.77e^{-06}$ - $2.17e^{-05}$
Upper South America	Dengue	1.04	1.10	$8.61e^{-01}$ - 1.26	1.25	1.13	$9.84e^{-01}$ - 1.58	$1.04e^{-04}$	$8.03e^{-05}$ - $1.32e^{-04}$

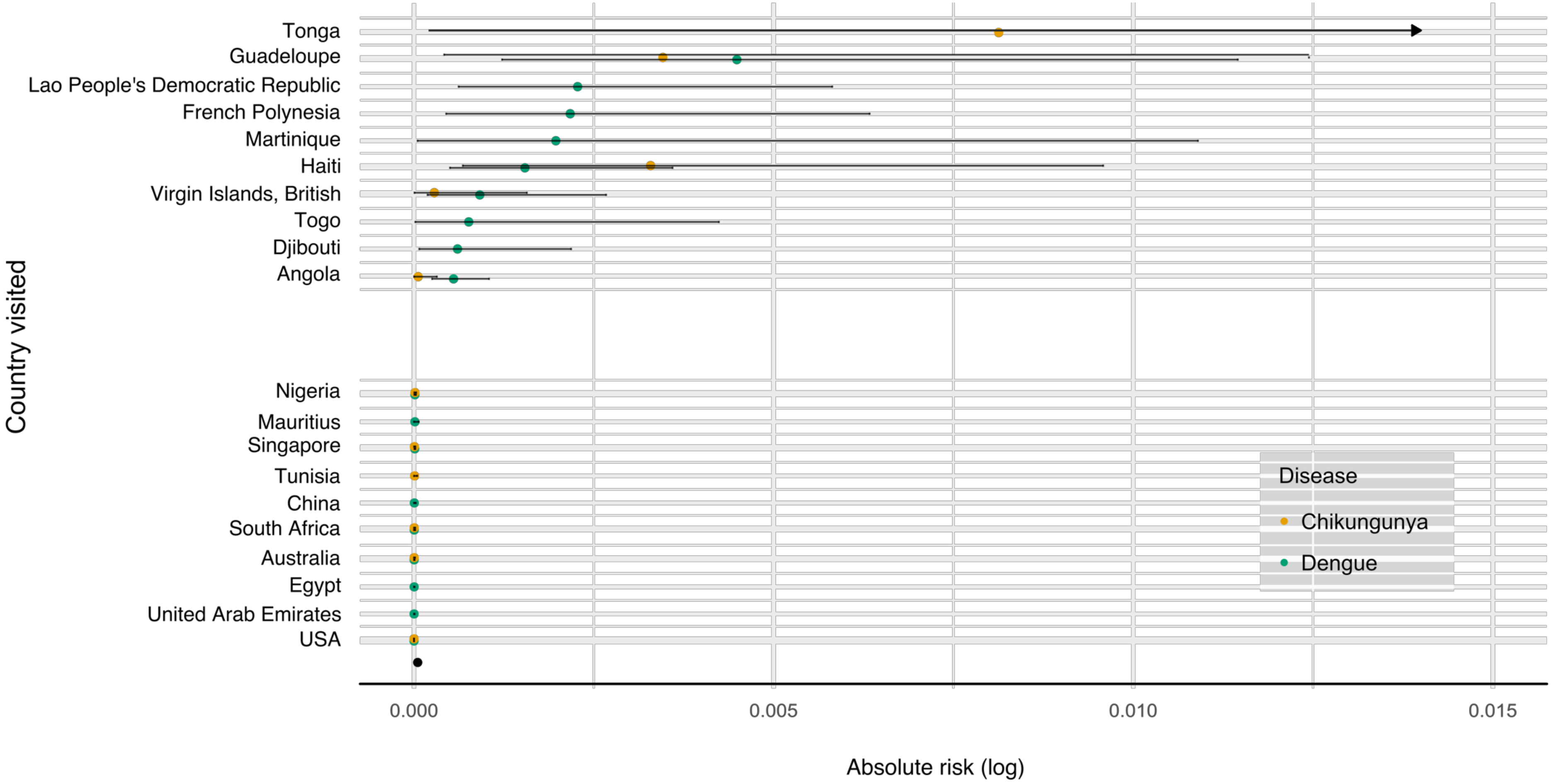


Figure 5.4: Absolute risk (imported cases / returning passengers) per country visited by chikungunya and dengue cases reported to PHE, restricted to the ten countries with the highest and lowest absolute risk values.

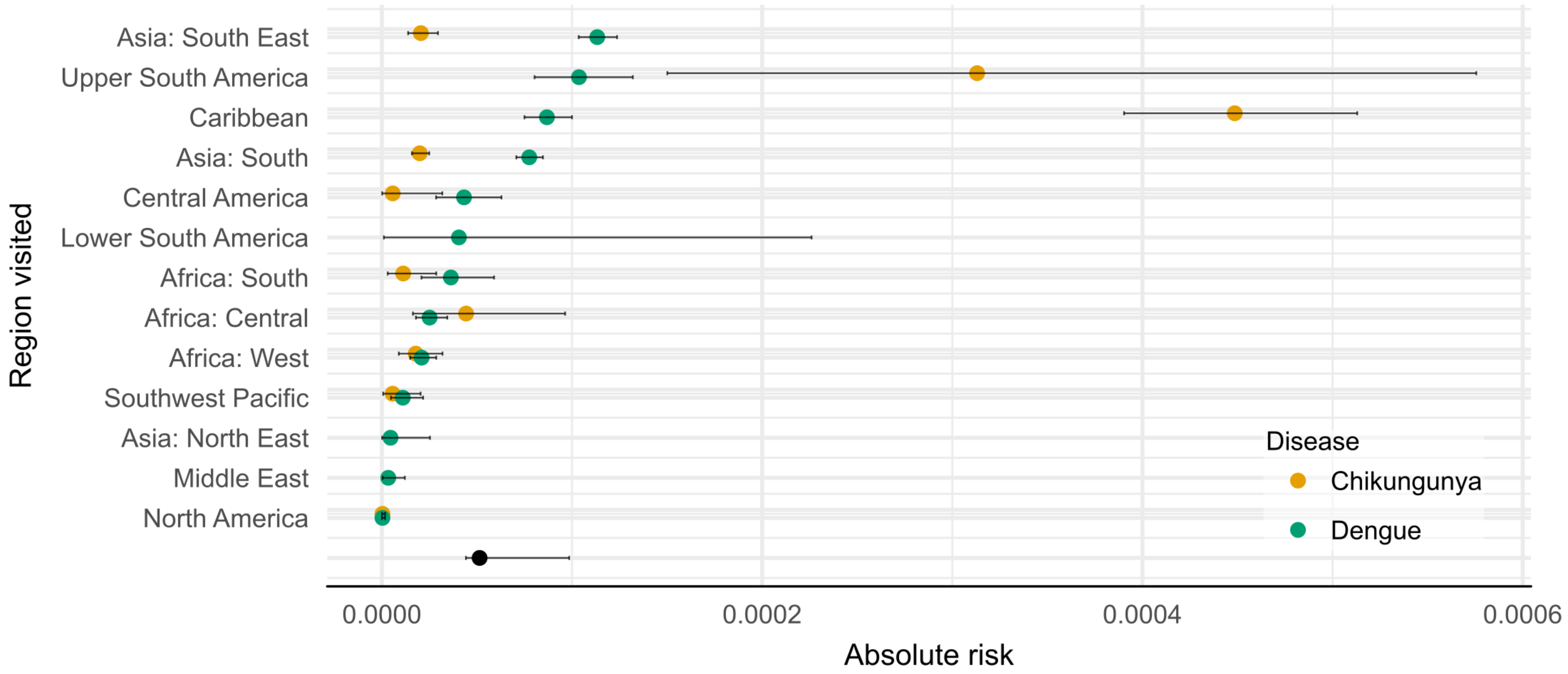


Figure 5.5: Absolulte risk (imported cases/returning passengers) per region (grouped neighbouring countries) visited by chikungunya and dengue cases reported to PHE.

Discussion

The annual number of chikungunya and dengue cases imported into the United Kingdom for the period ranging 2010 to 2014 was modelled as a function of the number of travellers and the known prevalence of each disease in the visited regions.

The results showed a seasonal and increasing trend in the number of detected cases imported in the UK from international travel. The largest number of imports occurred between August and November each year, whereas February saw the smallest number of imported cases. This trend reflects the seasonal patterns of both diseases in countries and regions in the Northern hemisphere (Johansson, 2015; San Martin *et al.*, 2010). The importation trend does not closely follow the number of airline bookings returning to the UK. Additionally, the structure of the age pyramid showed that younger age groups were more affected by dengue whereas older age groups by chikungunya, with the exception of the 35-39 year olds for chikungunya.

Understanding the passenger's demographic characteristics provides insights into their exposure risks. In the data shown here, there is a notable difference in the number of cases recorded for each disease when grouped by sex. Of the chikungunya cases, more females (58%) than males (42%) were diagnosed, whereas more male cases (55%) than female cases (44%) were diagnosed with dengue (with a small number of 'unknown' in this variable). Variations in the age groups affected by either disease were seen, with chikungunya affecting older generations (55-59 years old) and dengue younger generations (30-34 years).

The patterns were broadly consistent between regional-scale relative risk (regional coefficients) of infection for both diseases, and a range of regional-level coefficients. Regions where UK travellers had a lower relative risk of infection compared to the local populations, without considering duration of travel, were the Caribbean for both diseases and Upper South America for dengue only. However, regions such as Africa Central and Lower South America posed a higher relative risk to UK travellers relative to Caribbean with point estimates of relative risk suggestive of UK travellers being at higher relative risk than the local populations for dengue. It is worth noting that without considering duration of travel, North America and Southwest Pacific countries were associated with an increased risk of chikungunya virus infection in travellers compared to the local population, relative to the Caribbean. It must be remembered that Chikungunya was not present in the Americas before late 2013 and therefore the local population was in a naïve state of immunity. However, UK passengers to South East Asia were likely to have been at a reduced risk due to their within-country behaviour, affecting their exposure risk, which could not be modelled. The variations in disease reporting varied greatly by country and region and are likely to have influenced their relative risks.

Although not present in the original data, the impact of duration of travel on relative risk was modelled using imputed third-party data from the UK's Office for National Statistics. Including this data showed an overall protective effect by a factor of ten ($1.60e^{+3}$ compared to $4.60e^{+30}$) for passengers to all regions, with variations when considering each region in turn. For example, duration of travel showed a protective effect on chikungunya infection for passengers travelling to North America and Africa Central, but the opposite was true for passengers travelling to Southwest Pacific and Asia South. Regarding dengue infections, considering duration of travel had a protective effect for passengers travelling to Lower South America and Southwest Pacific but the opposite effect for those travelling to the Caribbean and Asia South East. The reasoning behind these variations was unclear, especially as these pathogens are transmitted by the same vectors. It is likely that these variations are a result of within-country behaviour and how much information and information passengers receive prior to travel, impacting their behaviour and whether precautions were taken prior or during travel.

The overall absolute risk encountered by UK passengers was low and in good agreement between diseases, except for a number of countries and regions, such as Guadeloupe and the Caribbean. In terms of airline bookings returning to the UK with chikungunya, the highest proportion returned from the Caribbean, especially the Guadeloupe, and those with dengue from South East Asia, especially Laos. However, the number of airline bookings to these regions and countries were relatively small.

Even though the largest number of cases in the data was reported from the Caribbean and South East Asia for chikungunya and dengue respectively, UK passengers travelling to these two regions were at reduced relative risk of infection compared to local populations, according to our model, before including duration of travel. Once duration was included, the Caribbean remained a safe destination in terms of dengue infection whereas South East Asia was a risky destination for chikungunya. On the other hand, countries in Upper South America were visited less frequently by chikungunya cases but presented a higher relative risk for travellers. The trend reflects the introduction of chikungunya in the Americas in December 2013, starting a large outbreak which affected over 43 countries, and caused a reported 1.4 million cases and 191 deaths (World Health Organization, 2017a). In the time period covered by this cases data set, a dengue outbreak was seen in the Portuguese island of Madeira in 2012, with a total of 23 known imported cases in the UK (European Centre for Disease Prevention and Control, 2013). Other large dengue outbreaks have not been recorded in Europe to this scale and/or during this time period.

The model developed here shows a good agreement between the regions with known outbreaks between 2010 and 2014, and their relative risk posed to UK travellers, even with known limitations. Regions affected by dengue outbreaks over the time period considered included South East Asia, Pacific Islands, Central America as well as Florida (United States) and China (World Health Organization, 2016a). However, regional results showed very large confidence intervals, which may be a result of localised outbreaks occurring within the associated countries, or little information available from the endemic countries. Indeed, the island of Madeira (Portugal, Western Africa for the purposes of this analysis) and the Yunnan region (China, North East Asia) saw localised dengue outbreaks, affecting a small percentage of the populations respectively (World Health Organization, 2016a). North America is very likely to have been impacted by the outbreaks seen in the rest of the continent, also affecting the size of its confidence intervals. However, India (South Asia) reported a large number of cases of chikungunya between 2010 and 2014 (ranging between 12,700 and 20,400 cases per year), across the country (Government of India, 2015). Such numbers are likely to have an impact on the model regional results. The variation in relative risks between both diseases for some regions such as Asia South and Central America, is likely to be a result of the varying levels of reporting for either disease. Indeed, as previously mentioned, chikungunya reporting is done less frequently than dengue in many countries, making estimating the level of endemic disease difficult. Finally, the countries visited by zika cases diagnosed by PHE (2014 to 2016 included) overlap the regions visited by dengue and chikungunya cases, further strengthening the need to develop a model suitable for all three VBDs.

Even if travellers are at reduced relative risk compared to local populations, personal protection must still be effectively taken to avoid illness. International travel advice providers such as NaTHNaC (travelhealthpro.org.uk) and the Centres for Disease Control and Prevention (wwwnc.cdc.gov/travel) both provide useful information to travellers regarding known countries at risk and prevention methods on a range of diseases that can be easily accessed. This information has implications in educating public health and point of care clinicians about pre-travel health advice, including information regarding personal protection methods, and post-travel infection treatment options in returning passengers.

Limitations

A number of limitations were known to the author when analysing this data, regarding both the data and the model.

Firstly, the relative risks of the two VBDs described here, at a specific travel destination may be difficult to determine, as the global burden of chikungunya and dengue can only be estimated based on the data freely available and are sometimes lacking in some countries where the exposure risk may be high (Nsoesie, 2016; World Health Organization, 2016a). Precisely determining the global burden of chikungunya is difficult given the varying levels of reporting provided by national and international health agencies (Nsoesie, 2016). These levels of reporting are also likely to change during the course of an epidemic, with suspected cases being reported at the start of an outbreak, followed by confirmed cases. This may give a false sense that cases numbers are decreasing when it is not the case. The model developed by Nsoesie (2016) suggests that 1.3 billion people live in areas at risk of chikungunya infection, whereas the WHO does not provide a global account of people living in at risk areas (Nsoesie, 2016). Even with such high numbers, chikungunya is an under-estimated and under-recognised problem globally, but especially in Africa, given the low mortality rates and misdiagnosis with dengue (Nsoesie, 2016). Although a potentially more severe disease given the possibility of haemorrhage from multiple dengue virus infections, dengue fever is also under-reported, and often misdiagnosed (Nsoesie, 2016). It is also estimated that 70% of dengue cases do not seek medical help when infected (Stanaway, 2016), and misdiagnosis may also contribute to under reporting (Nsoesie, 2016). Serological analyses would need to be done in countries with limited data availability to establish the true number of populations living in at-risk locations (Nsoesie, 2016), which may impact (positively or negatively) the model results. The endemic prevalence levels were calculated from the annual case numbers reported by each country, preventing any analysis of the seasonal and within-country variations. The absolute risk (returning cases among returning airline passenger bookings) is also likely to be underestimated given the potentially high number of asymptomatic cases as well as misdiagnosis by healthcare personnel. The original case data set only recorded known chikungunya and dengue cases in the United Kingdom, so travellers who were ill while abroad or who only experience mild (or asymptomatic) disease are unlikely to have been recorded.

The case data were not refined enough to provide the within-country destinations visited. This is of importance when looking at countries such as China and the United States with important environmental and climatic variations between northern and southern within-country regions. These variations determine vector habitat suitability and therefore potential outbreaks, as was seen in the chikungunya outbreak in the Guangdong province of southern China in 2010 (Nsoesie, 2016). Passenger’s purpose of travel was not recorded in the data either but will impact a passenger’s behaviour and therefore relative risk within the visited country. However, it is difficult to determine the impact this variables would have with certainty without any additional data; it was therefore assumed that all travellers behave in the same manner within and between countries. Appropriate knowledge of infection risks in visited counties may influence travellers into taking personal precautions, such as wearing appropriate clothing, minimising time spent outdoors at high risk times, and using insect repellent (World Health Organization, 2012). Tourists are more likely to behave differently to the local population by choosing to sleep in air conditioned hotels and/or only frequent locations that reduce their risk of infection (World Health Organization, 2012; Schlangenhauf, 2011). Therefore, understanding risk perceptions by traveller purpose of travel may help refine this analysis, as exposure risks are likely to change.

Lastly, it was assumed that all cases recorded in the imported case data started their journey in the UK and became infected while abroad. The equivalent OAG airline data was therefore selected to match this returning leg of a round trip, originating and ending in the UK, but flying back from an international destination. This excludes cases diagnosed by the National Health Service but who were residents of another country (unknown number).

Table 5.2: List of sources used to collect disease and population data for the regional risk with chikungunya and dengue among travellers returning to the UK. Note that the list of coutries reflects those visited by returning UK passengers cases during the time period considered (2010-2014).

Region name	Country name	Sources (chikungunya virus) (years included)	Sources (dengue virus) (years included)	Population (years included)
Africa: Central	Cameroon			World Bank Data (2010-2014) (data.worldbank.org/country/cameroon)
	Congo	International Federation Red Cross and Crescent Societies (2011) (www.ifrc.org/docs/Appeals/11/MDRCG007FR.pdf)		World Bank Data (2010-2014) (data.worldbank.org/country/congo-rep)
	Congo Democratic Republic of	World Health Organization (Epidemic and Pandemic Alert and Response Programme - Outbreak Report; July 25, 2011)		World Bank Data (2010-2014) (data.worldbank.org/country/congo-dem-rep)
	Djibouti			World Bank Data (2010-2014) (data.worldbank.org/country/djibouti)
	Egypt			World Bank Data (2010-2014) (data.worldbank.org/country/egypt)
	Eritrea			National Accounts Main Aggregates Database (unstats.un.org/unsd/snaama/dnllist.asp)
	Kenya		World Health Organization Africa, Regional Office Outbreak Bulletin (vol1, issue 6,Nov 2011) (2011) (www.afro.who.int/en/disease-outbreaks/outbreak-news/4155-dengue-outbreak-in-the-united-republic-of-tanzania-30-may-2014.html) (2014)	World Bank Data (2010-2014) (data.worldbank.org/country/kenya) (2010-2014)
	Rwanda			World Bank Data (2010-2014) (data.worldbank.org/country/rwanda)
	Somalia			World Bank Data (2010-2014) (data.worldbank.org/country/somalia)
	Sudan		World Health Organization regional office for the Eastern Mediterranean (applications.emro.who.int/dsaf/epi/2014/Epi_Monitor_2014_7_25.pdf?ua=1) (2010-2012) World Health Organization (2013) (reliefweb.int/sites/reliefweb.int/files/resources/Sudan%20Health%20Highlights%20Weeks%2022-23%20of%202013.pdf) European Centre for Disease Prevention and Control (ecdc.europa.eu/en/publications/publications/communicable-disease-threats-report-6-dec-2014.pdf) (2014)	World Bank Data (2010-2014) (data.worldbank.org/country/sudan)
	Tanzania		World Health Organization Africa, Regional Office (www.afro.who.int/en/disease-outbreaks/outbreak-news/4155-dengue-outbreak-in-the-united-republic-of-tanzania-30-may-2014.html) (2014)	World Bank Data (2010-2014) (data.worldbank.org/country/tanzania)
	Uganda			World Bank Data (2010-2014) (data.worldbank.org/country/uganda)
Southern Africa	Angola			World Bank Data (2010-2014) (data.worldbank.org/country/angola)
	Malawi			World Bank Data (2010-2014) (data.worldbank.org/country/malawi)
	Mauritius	Health Statistics Unit, Ministry of Health and Quality of Life (2012-2014) (health.govmu.org)		World Bank Data (2010-2014) (data.worldbank.org/country/mauritius)
	Mozambique			World Bank Data (2010-2014) (data.worldbank.org/country/mozambique)

(Table 5.2 continues on next page)

(Table 5.2 continued)

Region name	Country name	Sources (chikungunya virus) (years included)	Sources (dengue virus) (years included)	Population (years included)
Southern Africa	South Africa			World Bank Data (2010-2014) (data.worldbank.org/country/south-africa)
	Zambia			World Bank Data (2010-2014) (data.worldbank.org/country/zambia)
	Zimbabwe			World Bank Data (2010-2014) (data.worldbank.org/country/zimbabwe)
Africa West	Cote d'Ivoire			World Bank Data (2010-2014) (data.worldbank.org/country/cote-divoire)
	Gambia			World Bank Data (2010-2014) (data.worldbank.org/country/gambia)
	Ghana			World Bank Data (2010-2014) (data.worldbank.org/country/ghana)
	Guinea			World Bank Data (2010-2014) (data.worldbank.org/country/guinea)
	Liberia			World Bank Data (2010-2014) (data.worldbank.org/country/liberia)
	Nigeria			World Bank Data (2010-2014) (data.worldbank.org/country/nigeria)
	Madeira (Portugal)		European Centre for Disease Prenvention and Control (2010-2014) (ecdc.europa.eu/en/threats-and-outbreaks/reports-and-data/weekly-threats)	Direcao regional de estatica da Madeira (2010-2014) (estatistica.madeira.gov.pt/en/download-now-3/social-gb/popcondsoc-gb/demografia-gb/demografia-serie-gb/demografia-long-series-gb.html)
	Senegal			World Bank Data (2010-2014) (data.worldbank.org/country/senegal)
	Sierra Leone			World Bank Data (2010-2014) (data.worldbank.org/country/sierra-leone)
	Togo			World Bank Data (2010-2014) (data.worldbank.org/country/togo)
	Tunisia			World Bank Data (2010-2014) (data.worldbank.org/country/tunisia)
Asia North East	China	Qiaoli Z et al, (2012) (journals.plos.org/plosone/article?id=10.1371/journal.pone.0042830)	National Health and Family Planning Commission of the PRC (2012-2013) (en.nhfpc.gov.cn/2014-07/15/c_46865_2.htm) World Health Organization (2014) (www.wpro.who.int/emerging_diseases/dengue_biweekly_29dec2014.pdf?ua=1)	World Bank Data (2010-2014) (data.worldbank.org/country/china)
	Japan		Japan National Institute of Infectious Diseases (2010-2014) (www.niid.go.jp/niid/en/survei/2085-idwr/ydata/6058-report-ea2014-20.html)	World Bank Data (2010-2014) (data.worldbank.org/country/japan)
	Republic of Korea			World Bank Data (2010-2014) (data.worldbank.org/country/korea-rep)

(Table 5.2 continues on next page)

(Table 5.2 continued)

Region name	Country name	Sources (chikungunya virus) (years included)	Sources (dengue virus) (years included)	Population (years included)
South Asia	Afghanistan			World Bank Data (2010-2014) (data.worldbank.org/country/afghanistan)
	Bangladesh	Diop *et al* (2015) medcraveonline.com/MOJPH/MOJPH-02-00043.pdf (2013)	Government of the People's Republic of Bangladesh (2010-2014) (www.dghs.gov.bd/images/docs/Publicaations/HB%202015_1st_edition_311 22015.pdf)	World Bank Data (2010-2014) (data.worldbank.org/country/bangladesh)
	India	Government of India - national vector borne disease control programme, annual report 2014-15 (2010-2014) (nvbdcp.gov.in/doc/annual-report-nvbdcp-2014-15.pdf)	Government of India (2010-2014) (nvbdcp.gov.in/den-cd.html)	World Bank Data (2010-2014) (data.worldbank.org/country/india)
	Maldives	Institut de Veille Sanitaire – Bilan épidémiologique, Monde (invs.santepubliquefrance.fr//Dossiers-thematiques/Maladies-infectieuses/Maladies-a-transmission-vectorielle/Chikungunya/Donnees-epidemiologiques/Monde) (2010-2014)	Institut de Veille Sanitaire Département international (2010) (opac.invs.sante.fr/doc_num.php?explnum_id=8622) Republic of Maldives Ministry of Heath (2010-2014) (health.egov.mv/publications/50_Maldives_Health_Profile_2016_D1%203rd%20May.pdf)	World Bank Data (2010-2014) (data.worldbank.org/country/maldives)
	Nepal		Government of Nepal Ministry of health (2010-2014) (dohs.gov.np/wp-content/uploads/2016/06/Annual_Report_FY_2071_72.pdf)	World Bank Data (2010-2014) (data.worldbank.org/country/nepal)
	Pakistan		World Health Organization Eastern Mediterranean (2010-2014) (applications.emro.who.int/dsaf/epi/2013/Epi_Monitor_2013_6_37.pdf?ua=1)	World Bank Data (2010-2014) (data.worldbank.org/country/pakistan)
	Sri Lanka		Sri Lanka Ministry of Health (2010-2014) (www.epid.gov.lk/web/index.php?option=com_casesanddeaths&Itemid=448&lang=en#)	World Bank Data (2010-2014) (data.worldbank.org/country/sri-lanka)
South East Asia	Brunei Darussalam			World Bank Data (2010-2014) (data.worldbank.org/country/brunei-darussalam)
	Cambodia		World Health Organization Western Pacific Region (2010-2014) (http://www.wpro.who.int/emerging_diseases/DengueSituationUpdates/en/)	World Bank Data (2010-2014) (data.worldbank.org/country/cambodia)
	Chinese Taipei	Taiwan National Infectious Disease Statistics System (2010-2014) (nidss.cdc.gov.tw/en/CDCWNH07.aspx?dc=1&dt=2&disease=A920)	Centre for Disease Control, ROC (Taiwan) (2010-2014) (www.cdc.gov.tw/english/info.aspx?treeid=e79c7a9e1e9b1cdf&nowtreeid=e02c24f0dacdd729&tid=D76AD76D26365478)	National Statistics Republic of China (2010-2014) (eng.stat.gov.tw/lp.asp?CtNode=6339&CtUnit=1072&BaseDSD=36&mp=5)
	Hong Kong		Government of the Hong Kong Special Administrative Region - Department of health (2010-2014) (www.chp.gov.hk/en/data/1/10/26/43/2285.html)	World Bank Data (2010-2014) (data.worldbank.org/country/hong-kong-sar-china)
	Indonesia	Ministry of Health Republic of Indonesia - Indonesia Health profile 2014 (2010-2014) (www.depkes.go.id/resources/download/pusdatin/profil-kesehatan-indonesia/Indonesia%20Health%20Profile%202014.pdf)	World Health Organization South East Asia (2010-2012) (www.searo.who.int/entity/vector_borne_tropical_diseases/data/graphs.pdf?ua=1)	World Bank Data (2010-2014) (data.worldbank.org/country/indonesia)
	Lao		World Health Organization Western Pacific Region (2010-2014) (www.wpro.who.int/emerging_diseases/dengue_biweekly_20160113.pdf?ua=1)	World Bank Data (data.worldbank.org/country/lao-pdr) (2010-2014)

(Table 5.2 continues on next page)

(Table 5.2 continued)

Region name	Country name	Sources (chikungunya virus) (years included)	Sources (dengue virus) (years included)	Population (years included)
		Institut de Veille Sanitaire – Bilan épidémiologique, Monde (invs.santepubliquefrance.fr//Dossiers-thematiques/Maladies-infectieuses/Maladies-a-transmission-vectorielle/Chikungunya/Donnees-epidemiologiques/Monde) (2010-2012)	World Health Organization Western Pacific Region (2010-2014) (http://www.wpro.who.int/emerging_diseases/documents/Dengue_Archives/en/)	World Bank Data (2010-2014) (data.worldbank.org/country/malaysia)
	Myanmar		World Health Organization South East Asia Region (2010-2012) (www.searo.who.int/entity/vector_borne_tropical_diseases/data/graphs.pdf?ua=1) Myanmar Ministry of Information (2013-2014) (www.moi.gov.mm/moi:eng/?q=news/29/06/2015/id-4205)	World Bank Data (2010-2014) (data.worldbank.org/country/myanmar)
	Philippines		World Health Organization South East Asia Region (2010-2014) (www.wpro.who.int/emerging_diseases/documents/Dengue_Archives/en/)	World Bank Data (2010-2014) (data.worldbank.org/country/philippines)
South East Asia	Singapore	Ministry of Health Singapore (2010-2014) (www.moh.gov.sg/content/moh_web/home/Publications/Reports/)	World Health Organization South East Asia Region (2010-2014) (www.wpro.who.int/emerging_diseases/documents/Dengue_Archives/en/)	World Bank Data (2010-2014) (data.worldbank.org/country/singapore)
	Thailand	Thailand Ministry of Health (2010-2014) (www.boe.moph.go.th/Annual/AESR2014/aesr2557/Part%202/table05.pdf)	Thailand Ministry of Health (2010-2014) (www.boe.moph.go.th/Annual/AESR2014/aesr2557/Part%202/table05.pdf)	World Bank Data (2010-2014) (data.worldbank.org/country/thailand)
	Timor Leste			World Bank Data (2010-2014) (data.worldbank.org/country/timor-leste)
	Viet Nam		World Health Organization South East Asia Region (2010-2014) (www.wpro.who.int/emerging_diseases/documents/Dengue_Archives/en/)	World Bank Data (2010-2014) (data.worldbank.org/country/vietnam)
	Antigua and Barbuda, Leeward Islands	Pan American Health Organisation (PAHO) (2013-2014) (www.paho.org/hq/index.php?option=com_topics&view=readall&cid=5927&Itemid=40931&lang=en)	Pan American Health Organisation (PAHO) (2010-2014) (www.paho.org/hq/index.php?option=com_docman&task=doc_view&gi d=23999&Itemid=&lang=en)	World Bank Data (2010-2014) (data.worldbank.org/country/antigua-and-barbuda)
	Barbados	Pan American Health Organisation (PAHO) (2013-2014) (www.paho.org/hq/index.php?option=com_topics&view=readall&cid=5927&Itemid=40931&lang=en)	Pan American Health Organisation (PAHO) (2010-2014) (www.paho.org/hq/index.php?option=com_docman&task=doc_view&gid=23999&Itemid=&lang=en)	World Bank Data (2010-2014) (data.worldbank.org/country/barbados)
	Cuba		Pan American Health Organisation (PAHO) (2010-2014) (www.paho.org/hq/index.php?option=com_docman&task=doc_view&gid=23999&Itemid=&lang=en)	World Bank Data (2010-2014) (data.worldbank.org/country/cuba)
Caribbean	Dominica	Pan American Health Organisation (PAHO) (2013-2014) (www.paho.org/hq/index.php?option=com_topics&view=readall&cid=5927&Itemid=40931&lang=en)	Pan American Health Organisation (PAHO) (2010-2014) (www.paho.org/hq/index.php?option=com_docman&task=doc_view&gid=23999&Itemid=&lang=en)	World Bank Data (2010-2014) (data.worldbank.org/country/dominica)
	Dominican Republic	Pan American Health Organisation (PAHO) (2013-2014) (www.paho.org/hq/index.php?option=com_topics&view=readall&cid =5927&Itemid=40931&lang=en)	Pan American Health Organisation (PAHO) (2010-2014) (www.paho.org/hq/index.php?option=com_docman&task=doc_view&gid=23999& Itemid=&lang=en)	World Bank Data (2010-2014) (data.worldbank.org/country/dominican-republic)
	Grenada	Pan American Health Organisation (PAHO) (2013-2014) (www.paho.org/hq/index.php?option=com_topics&view=readall&cid=5927&Itemid=40931&lang=en)	Pan American Health Organisation (PAHO) (2010-2014) (www.paho.org/hq/index.php?option=com_docman&task=doc_view&gid=23999&Itemid=&lang=en)	World Bank Data (2010-2014) (data.worldbank.org/country/grenada)

(Table 5.2 continues on next page)

(Table 5.2 continued)

Region name	Country name	Sources (chikungunya virus) (years included)	Sources (dengue virus) (years included)	Population (years included)
Caribbean	Guadeloupe	Pan American Health Organisation (PAHO) (2013-2014) (www.paho.org/hq/index.php?option=com_topics&view=readall&cid=5927&It emid=40931&lang=en)	Pan American Health Organisation (PAHO) (2010-2014) (www.paho.org/hq/index.php?option=com_docman&task=doc_view&gid=23999&Itemid=&lang=en)	Institut d'Emission Des Departements d'Outre-Mer (2010-2014) (www.iedom.fr/IMG/pdf/ra_2015_iedom_gua.pdf)
	Haiti	Pan American Health Organisation (PAHO) (2013-2014) (www.paho.org/hq/index.php?option=com_topics&view=readall&cid=5927&Itemid=40931&lang=en)	Pan American Health Organisation (PAHO) (2010-2014) (www.paho.org/hq/index.php?option=com_docman&task=doc_view&gid=23999&Itemid=&lang=en)	World Bank Data (2010-2014) (data.worldbank.org/country/haiti)
	Jamaica	Pan American Health Organisation (PAHO) (2013-2014) (www.paho.org/hq/index.php?option=com_topics&view=readall&cid=5927&Itemid=40931&lang=en)	Pan American Health Organisation (PAHO) (2010-2014) (www.paho.org/hq/index.php?option=com_docman&task=doc_view&gid=23999&Itemid=&lang=en)	World Bank Data (2010-2014) (data.worldbank.org/country/jamaica)
	Martinique		Pan American Health Organisation (PAHO) (2010-2014) (www.paho.org/hq/index.php?option=com_docman&task=doc_view&gid=23999&Itemid=&lang=en)	Institut national de la statistique et des etudes economiques (2010-2013) (www.insee.fr/fr/accueil) Institut d'Emission d'Outre-Mer (2014) (www.iedom.fr/IMG/pdf/ra_2015_iedom_mar_.pdf)
	Netherlands Antilles		Pan American Health Organisation (PAHO) (2010-2014) (www.paho.org/hq/index.php?option=com_docman&task=doc_view&gid=23999&Itemid=&lang=en)	Statistics Netherlands (2010-2014) (statline.cbs.nl/Statweb/publication/?VW=T&DM=SLEN&PA=80534ENG&D1=0&D2=0&D3=0&D4=a&D5=8-13&HD=160721-1631&LA=EN&HDR=T,G2,G1,G4&STB=G3)
	Puerto Rico	Pan American Health Organisation (PAHO) (2013-2014) (www.paho.org/hq/index.php?option=com_topics&view=readall&cid=5927&Itemid=40931&lang=en)	Pan American Health Organisation (PAHO) (2010-2014) (www.paho.org/hq/index.php?option=com_docman&task=doc_view&gid=23999&Itemid=&lang=en)	World Bank Data (2010-2014) (data.worldbank.org/country/puerto-rico)
	Saint Lucia		Pan American Health Organisation (PAHO) (2010-2014) (www.paho.org/hq/index.php?option=com_docman&task=doc_view&gid=23999&Itemid=&lang=en)	World Bank Data (2010-2014) (data.worldbank.org/country/st-lucia)
	Saint Martin		Pan American Health Organisation (PAHO) (2010-2014) (www.paho.org/hq/index.php?option=com_docman&task=doc_view&gid=23999&Itemid=&lang=en)	World Bank Data (2010-2014) (data.worldbank.org/country/st-martin-french-part)
	St Vincent and the Grenadines	Pan American Health Organisation (PAHO) (2013-2014) (www.paho.org/hq/index.php?option=com_topics&view=readall&cid=5927&Itemid=40931&lang=en)	Pan American Health Organisation (PAHO) (2010-2014) (www.paho.org/hq/index.php?option=com_docman&task=doc_view&gid=23999&Itemid=&lang=en)	World Bank Data (2010-2014) (data.worldbank.org/country/st-vincent-and-the-grenadines)
	Trinidad and Tobago	Pan American Health Organisation (PAHO) (2013-2014) (www.paho.org/hq/index.php?option=com_topics&view=readall&cid=5927&Itemid=40931&lang=en)	Pan American Health Organisation (PAHO) (2010-2014) (www.paho.org/hq/index.php?option=com_docman&task=doc_view&gid=23999&Itemid=&lang=en)	World Bank Data (2010-2014) (data.worldbank.org/country/trinidad-and-tobago)
	Virgin Islands, British	Pan American Health Organisation (PAHO) (2013-2014) (www.paho.org/hq/index.php?option=com_topics&view=readall&cid=5927&Itemid=40931&lang=en)	Pan American Health Organisation (PAHO) (2010-2014) (www.paho.org/hq/index.php?option=com_docman&task=doc_view&gid=23999&Itemid=&lang=en)	World Bank Data (2010-2014) (data.worldbank.org/country/british-virgin-islands)
Central America	Costa Rica		Pan American Health Organisation (PAHO) (2010-2014) (www.paho.org/hq/index.php?option=com_docman&task=doc_view&gid=23999&Itemid=&lang=en)	World Bank Data (2010-2014) (data.worldbank.org/country/costa-rica)
	El Salvador		Pan American Health Organisation (PAHO) (2010-2014) (www.paho.org/hq/index.php?option=com_docman&task=doc_view&gid=23999&Itemid=&lang=en)	World Bank Data (2010-2014) (data.worldbank.org/country/el-salvador)

(Table 5.2 continues on next page)

(Table 5.2 continued)

Region name	Country name	Sources (chikungunya virus) (years included)	Sources (dengue virus) (years included)	Population (years included)
Central America	Guatemala	Pan American Health Organisation (PAHO) (2014) (www.paho.org/hq/index.php?option=com_topics&view=readall&cid=5927&Itemid=40931&lang=en)	Pan American Health Organisation (PAHO) (2010-2014) (www.paho.org/hq/index.php?option=com_docman&task=doc_view&gid=23999&Itemid=&lang=en)	World Bank Data (2010-2014) (data.worldbank.org/country/guatemala)
	Honduras		Pan American Health Organisation (PAHO) (2010-2014) (www.paho.org/hq/index.php?option=com_docman&task=doc_view&gid=23999&Itemid=&lang=en)	World Bank Data (2010-2014) (data.worldbank.org/country/honduras)
	Mexico	Pan American Health Organisation (PAHO) (2014) (www.paho.org/hq/index.php?option=com_topics&view=readall&cid=5927&Itemid=40931&lang=en)	Pan American Health Organisation (PAHO) (2010-2014) (www.paho.org/hq/index.php?option=com_docman&task=doc_view&gid=23999&Itemid=&lang=en)	World Bank Data (2010-2014) (data.worldbank.org/country/mexico)
	Nicaragua		Pan American Health Organisation (PAHO) (2010-2014) (www.paho.org/hq/index.php?option=com_docman&task=doc_view&gid=23999&Itemid=&lang=en)	World Bank Data (2010-2014) (data.worldbank.org/country/nicaragua)
	Panama		Pan American Health Organisation (PAHO) (2010-2014) (www.paho.org/hq/index.php?option=com_docman&task=doc_view&gid=23999&Itemid=&lang=en)	World Bank Data (2010-2014) (data.worldbank.org/country/panama)
Lower South America	Argentina		Pan American Health Organisation (PAHO) (2010-2014) (www.paho.org/hq/index.php?option=com_docman&task=doc_view&gid=23999&Itemid=&lang=en)	World Bank Data (2010-2014) (data.worldbank.org/country/argentina)
	Bolivia	Pan American Health Organisation (PAHO) (2014) (www.paho.org/hq/index.php?option=com_topics&view=readall&cid=5927&Itemid=40931&lang=en)		World Bank Data (2010-2014) (data.worldbank.org/country/bolivia)
	Peru		Pan American Health Organisation (PAHO) (2010-2014) (www.paho.org/hq/index.php?option=com_docman&task=doc_view&gid=23999&Itemid=&lang=en)	World Bank Data (2010-2014) (data.worldbank.org/country/peru)
Middle East	Iran			World Bank Data (2010-2014) (data.worldbank.org/country/iran-islamic-rep)
	United Arab Emirates			World Bank Data (2010-2014) (data.worldbank.org/country/united-arab-emirates)
	Saudi Arabia			World Bank Data (2010-2014) (data.worldbank.org/country/saudi-arabia)
	Yemen			World Bank Data (2010-2014) (data.worldbank.org/country/yemen)
North America	USA	Pan American Health Organisation (PAHO) (2014) (www.paho.org/hq/index.php?option=com_topics&view=readall&cid=5927&Itemid=40931&lang=en)	Pan American Health Organisation (PAHO) (2010-2014) (www.paho.org/hq/index.php?option=com_docman&task=doc_view&gid=23999&Itemid=&lang=en)	World Bank Data (2010-2014) (data.worldbank.org/country/united-states)
Southwest Pacific	Australia	Australian Government Department of Health - National Notifiable Diseases Surveillance System (2010-2014) (www9.health.gov.au/cda/source/rpt_2.cfm)		World Bank Data (2010-2014) (data.worldbank.org/country/australia)
	Fiji		Fiji Journal of Public Health (2010) (www.health.gov.fj/PDFs/Fiji%20Journal%20of%20Public%20Health%20Vol2Issue2.pdf) Fiji Ministry of health and medical services (2011-2014) (www.health.gov.fj/?page_id=198#1)	World Bank Data (2010-2014) (data.worldbank.org/country/fiji)

(Table 5.2 continues on next page)

(Table 5.2 continued)

Region name	Country name	Sources (chikungunya virus) (years included)	Sources (dengue virus) (years included)	Population (years included)
Southwest Pacific	French Polynesia	Institut Pasteur de Nouvelle Calédonie – Rapport sur les activités de sante publique (2011-2014) (www.institutpasteur.nc/wp-content/uploads/2014/07/Rapport-IPNC-Sant%C3%A9-Publique-2014-n%C2%B060-2015-du-30.3.15.pdf)		World Bank Data (2010-2014) (data.worldbank.org/country/french-polynesia)
	New Zealand		New Zealand Public Health Observatory (2010-2014) (www.nzpho.org.nz/NotifiableDisease.aspx)	World Bank Data (2010-2014) (data.worldbank.org/country/new-zealand)
	Papua New Guinea			World Bank Data (2010-2014) (data.worldbank.org/country/papua-new-guinea)
	Solomon Islands			World Bank Data (2010-2014) (data.worldbank.org/country/solomon-islands)
	Tonga			World Bank Data (2010-2014) (data.worldbank.org/country/tonga)
Upper South America	Brazil	Pan American Health Organisation (PAHO) (2014) (www.paho.org/hq/index.php?option=com_topics&view=readall&cid=5927&Itemid=40931&lang=en)	Pan American Health Organisation (PAHO) (2010-2014) (www.paho.org/hq/index.php?option=com_docman&task=doc_view&gi d=23999&Itemid=&lang=en)	World Bank Data (2010-2014) (data.worldbank.org/country/brazil)
	Colombia	Pan American Health Organisation (PAHO) (2014) (www.paho.org/hq/index.php?option=com_topics&view=readall&cid=5927&Itemid=40931&lang=en)		World Bank Data (2010-2014) (data.worldbank.org/country/colombia)
	Ecuador		Pan American Health Organisation (PAHO) (2010-2014) (http://www.paho.org/hq/index.php?option=com_docman&task=doc_view&gid=18233&Itemid=&lang=en)	World Bank Data (2010-2014) (data.worldbank.org/country/ecuador)
	French Guiana		Pan American Health Organisation (PAHO) (2010-2014) (www.paho.org/hq/index.php?option=com_docman&task=doc_view&gid=18233&Itemid=&lang=en)	Institut d'Emission Des Departements d'Outre-Mer (2010-2014) (www.iedom.fr/IMG/pdf/ra_2015_iedom_guy_pour_menl.pdf)
	Guyana	Pan American Health Organisation (PAHO) (2014) (www.paho.org/hq/index.php?option=com_topics&view=readall&cid=5927&Itemid=40931&lang=en)	Pan American Health Organisation (PAHO) (2010-2014) (www.paho.org/hq/index.php?option=com_docman&task=doc_view&gid=18233&Itemid=&lang=en)	World Bank Data (2010-2014) (data.worldbank.org/country/guyana)
	Venezuela	Pan American Health Organisation (PAHO) (www.paho.org/hq/index.php?option=com_topics&view=readall&cid=5927&Itemid=40931&lang=en)		World Bank Data (2010-2014) (data.worldbank.org/country/venezuela)

Chapter 6 – Understanding the possible origin of the next pandemic using airline travel patterns and healthcare development.

Preamble

There is an abundance of literature stating that weak healthcare systems are ideal settings for the emergence of an outbreak with slow within-country detection (Barber *et al.*, 2017; Bonds *et al.*, 2018; Elmahdawy *et al.*, 2017; Moore *et al.*, 2016). Although Bogoch *et al.* (2018) did relate the Madagascar healthcare system and the country's global connectivity in the context of the plague outbreak, no work has yet attempted, to the author's knowledge, to link national healthcare development and global connectivity to estimate the potential impact of such outbreaks for the global community. It is hoped that this novel approach will give new insights for future pandemic preparedness.

Abstract

Pandemics may spread very rapidly around the world and can have significant costs associated to them, in terms of economic and health impacts. Early detection of an outbreak is key to its control and to limiting its further spread, both nationally and internationally. Given the significant level of global connectivity described in **Chapter 3**, an uncontrolled outbreak in one country may quickly reach other countries and develop into a pandemic. The aim of this chapter was to explore variation in potential country-level risk for novel outbreaks to go undetected and for the pathogen to spread internationally via air-passenger travel.

The approach used in this chapter was to compare data regarding each country's healthcare development against its global airline connectivity, assigning equal weights to each. It was originally thought that a single parameter could be used as a proxy for healthcare development; however, the use of two indices made up of multiple parameters was ultimately deemed to be more appropriate. Global connectivity was estimated by generating an information flow matrix from the global airline data set downloaded from OAG. A fictitious 'worst case scenario' (WCS) country was assigned the best connectivity value of the network and the worst healthcare development score. Each country's relative proximity to WCS was subsequently calculated and plotted according to each index value.

The results indicate that India and Pakistan were the two closest countries to the WCS point for both indices, and were thereby postulated to pose the greatest risk to the global community. Additionally, countries that have recently seen the spread of outbreaks develop into pandemics (such as Brazil (Zika) and Mexico (H1N1), for example) were also identified as being relatively high potential threat to the global community. On the other hand, countries such as Monaco, Tuvalu and Slovenia were shown as posing the lowest risk.

This analysis highlights the importance of considering a country's connectivity as well as healthcare development when considering its potential impact in the spread of the next pandemic. In a world increasingly well connected, an outbreak in one country should be of concern for the global community. In order to reduce the global financial burden and reduce the mortality and morbidity, healthcare development and global connectivity should be considered together. This analysis highlights the potential risk posed to the global community of not detecting outbreaks early through strong healthcare systems. The international community could consider the potential benefits of additional support aimed towards those countries with the potential to cause the highest risk to the global community.

Introduction

In the wake of the Zika pandemic and Ebola outbreak, airline travel and level of healthcare provided by each country, have been demonstrated to impact a country's ability to contain outbreaks. The early detection and containment of an outbreak are of critical importance for the country's population health, as well as for the global community, to prevent a potential pandemic (International Working Group on Financing Preparedness, 2017). The Sustainable Development Goals (SDGs) require countries to improve their quality and access to healthcare (goal 3) and meet specific targets by 2030 (United Nations, 2017), as these are important tools to preventing and controlling future outbreaks. If a system is ready (well established with appropriate facilities, equipment and trained staff), it may be more likely to detect an outbreak quickly with trained staff triggering the appropriate response systems (Bonds *et al.*, 2018; Moore *et al.*, 2016). Although access to quality healthcare has greatly improved globally since 1990, the gap between countries with good or poor healthcare access is widening. Countries offering the best quality and access to healthcare are found in Western Europe, especially in Scandinavia, whereas the poorest quality and access is provided in Sub Saharan Africa and Oceania, according to the Global Burden of Disease's recent work (Barber *et al.*, 2017).

As described in previous chapters, the global airline network is growing at an accelerated pace, linking geographically distant countries (Glaesser *et al.*, 2017). As well as the increasing number of passengers, airline travel is continuously getting quicker, with the first Perth (Australia) to London (UK) direct flight landing in March 2018, linking the two countries in 17 hours (BBC, 2018). A non-negligible threat to global health today is the possibility of travelling to the other side of the world before becoming symptomatic, with passenger numbers and distances travelled both increasing rapidly (World Health Organization, 2018b). Infected passengers travelling within the airline network to epidemiologically suitable locations have the potential to propagate outbreaks through onward transmission (Tian *et al.*, 2017). Airline travel is the main access route to some remote locations (Bobashev *et al.*, 2008) and given the different exposure risks between local and visiting populations (Mier *et al.*, 2017), this is the most likely means of distant international disease spread (Bobashev *et al.*, 2008).

Some countries' airports are also increasingly being used as hubs (airports with a large number of connections to other airports) to reach other destinations where direct flights may be absent or rare (Wandelt and Sun, 2015). These country hubs include India, Singapore and Thailand, as the Asian air travel has seen a sharp growth in number of passengers and airports since the 1980s (Tian *et al.*, 2017). Although the overall network diameter is getting smaller (increasing number of flights and passengers to cross the entire network) (Huang *et al.*, 2013), as countries are increasingly well connected, some nations, especially island nations such as Papua New Guinea and Falkland Islands remain relatively hard to reach, and are therefore more distantly connected to the rest of the network (Guimera *et al.*, 2005).

A pandemic's dissemination from its source strongly depends on the local connectivity of the source country and the time for an outbreak to arrive through the airline network is independent of disease characteristics (Lawyer, 2016). Within the airline network, a total of 73 airports (2%) act as major hubs, with other airports acting as bridges between these hubs, peripheral airports (28%) and regional population centres. For example, following strong economic cooperation, the United Arab Emirates acts as a bridge between Europe and Asia, as well as between Asia and Africa (Wandelt and Sun, 2015). It must be noted that city size and their flight volumes are not closely related: large cities aren't always the most connected or have the largest passenger volumes (Bobashev *et al.*, 2008). However, a group of nodes (airports) can act as a clique by being well linked to each other, such that any node within a clique can be reached quickly in a few steps (Wandelt and Sun, 2015). Countries such as the United States and the United Kingdom, among others, are highly influential in terms of passenger transfers within the network, due to their population size and GDP (Wandelt and Sun, 2015).

Pandemics are unpredictable (geographically and temporally) and very costly, both in terms of financial losses, fatalities and morbidity; therefore, investing in preparedness at the country level should be encouraged. A country's pandemic preparedness (and response) relies heavily on the capacity of rapid detection and response to outbreaks. Accurate and timely detection by surveillance systems allows an outbreak to be identified quickly, with ongoing monitoring of cases assisting with the goal of avoiding important consequences both nationally and internationally, as was seen in the West African Ebola outbreak (International Working Group on Financing Preparedness, 2017). Although knowing or estimating the level of imported cases by travellers is important, it must also be noted that the risk may be underestimated if this is based on import notification alone (Lopez *et al.*, 2016). The 2014 West Africa Ebola outbreak is an example of the potential impact of an outbreak on an unstable healthcare system and *vice versa* (International Working Group on Financing Preparedness, 2017; Omoleke *et al.*, 2016). Although Guinea, Liberia and Sierra Leone had achieved significant economic progress through post-conflict reconstruction, access to good quality healthcare was still limited. Prior to the outbreak, a total of 39 inter-continental flights to and from these countries were operated weekly by major airlines, thus creating a potentially important risk of spreading the virus to the global community. This potential risk led to the suspension of flights to and from the region (Omoleke *et al.*, 2016), even though this was not advised by the WHO (Nutall, 2014). It was additionally noted that most sub-Saharan African countries did not have the capacities to contain the viral outbreak due to inadequate health facilities (Omoleke *et al.*, 2016).

Understanding the vulnerability of countries in the event of an outbreak has recently become a focus for some international groups, including the International Working Group on Financial Preparedness set up by the World Bank in November 2016, with the aim of identifying sectors which could help countries be more prepared for a pandemic (International Working Group on Financing Preparedness, 2017). Additionally, the WHO and World Bank have very recently co-created the Global Preparedness Monitoring Board with the goal of improving global preparedness (World health Organization, 2018e). Several groups have also developed pandemic preparedness indices to evaluate how vulnerable or resilient countries are to a potential pandemic (International Working Group on Financing Preparedness, 2017). In response to the Severe Acute Respiratory Syndrome (SARS) pandemic, the WHO updated its International Health Regulations (IHR) in 2005, requiring all member states to report in a timely manner any outbreak posing a possible international public health threat. However, few (64) countries had reported to WHO as having good surveillance systems in place, and 48 would not be capable to cope with a significant outbreak. Early detection and reporting remains crucial for a quick and effective response, ultimately costing fewer lives (International Working Group on Financing Preparedness, 2017).

The aim of this study was to further evaluate global preparedness by understanding how likely an outbreak would be to spread internationally when started from a given seed country. To do this, every country's global connectivity was estimated using a global network analysis of airline passenger bookings and pandemic preparedness levels estimated using a range of infectious disease proxies and health system indices.

Methods

Measure of global connectivity

To determine the level of connectivity between countries, a network percolation model was developed using global airline bookings. The airline data used for this model was the Traffic Analyser data set from OAG, spanning February 2010 to May 2015, downloaded between August 2014 and July 2015. The data represents the number of monthly bookings for routings between international airports, including any stop overs aggregated by country, both domestically and internationally. The percolation model simulates the spread of 'information' through a network, where country nodes exist in one of two states: empty or occupied. This model is analogous to an SI epidemic model, but should not be considered equivalent to a pandemic model as no within country transmission occurs, and the 'force of infection' represented by an occupied country remains constant over time. The probability that node i with empty status becomes occupied, p_i, is determined by the number of occupied nodes it is connected to, the rate of passenger flow along the network edges for a specific month, F, and a rate coefficient, β of value 1e^{-6} (arbitrary value):

$$p_i(t) = 1 - \exp\,(-\beta \sum_{j \in \Psi(\mathrm{t}), \mathrm{j} \neq \mathrm{i}} F_{ij})$$

where j denotes other nodes, and $\Psi(\mathrm{t})$ is the set of occupied nodes at time t. This model was implemented as a discrete time, stochastic Markov process, where all countries (nodes) are empty at the start of a simulation apart from a designated node, the seed country. Simulations proceeded until at least 25% of all nodes had been occupied (arbitrary values), at which point the simulation time was recorded (end time, ϕ). Simulations were performed 1,000 times for each seed country (243 in total) and month combination (64 in total) in turn. A cap of 2,000 iterations was imposed (arbitrary value), though in practise $\phi \ll 2000$. The mean and 2.5% and 97.5% percentiles of $\bar{\phi}_{i,m} = \frac{\sum_k \phi_{i,m,k}}{K}$ were calculated across K=1,000 replicates as a summary of the connectivity associated with each seed country and each month of airline data. Further average across all months and consider the connectivity as a rate such that the connectivity θ for country *i* was given by:

$$\theta_i = \frac{KM}{\sum_m \sum_k \phi_{i,m,k}}$$

To calculate **Figure 6.1** below the average across all countries was used and the connectivity was consider as a rate such that the connectivity by month for C the total number of countries is:

$$\theta_i = \frac{KC}{\sum_i \sum_k \phi_{i,m,k}}$$

Additionally, to calculate **Figure 6.2**, the same method was used as for **Figure 6.1** but month 55 only.

Understanding the level of healthcare provided

Although the structure of a healthcare system is complicated, involving many governmental branches, private and international organizations, it was initially hoped to use a single indicator as a proxy of the level of healthcare provided within each country. However, very limited information was available regarding which factors would be best to use as proxies, so five published indicators were considered (**Table 6.1**). It was necessary for the indicators to be relatable to all countries and not be specific to a given region (for example, malaria related indicators could not be applied to all countries given the parasite's geographic distribution) and to be indicators all countries measure routinely. Note that 'regions' in this context were defined as a group of neighbouring countries.

The WHO published a list of World Health Statistics Indicators (World Health Organization, 2015a), with a description of each indicator, from which the indicators used for this analysis were chosen. However, some drawbacks could be noted against the use of each one including the difficulty to dissociate the role of external funding from governmental. Each factor was used in turn to determine how likely a given country was to detect and control a novel infectious disease outbreak. All factors were plotted against each other and the Pearson correlation coefficients calculated, to determine possible associations between them (See **Appendix**, **Table 2**).

No strong correlation could be found between these indicators, with the strongest correlation coefficient of 0.63 between life expectancy and measles vaccination (95% confidence intervals ranging between 0.54 and 0.70). The absence of correlation between indicators implied that the indicators affected different parts of the healthcare within each country. Such variation suggested that using a single indicator was not reasonable and therefore a combination of several indicators may be needed to be used instead. These findings were supported by Moore *et al.* (2016), who also found that healthcare systems could not be adequately described by a single factor. Given the need for a more holistic understanding of healthcare systems, epidemic preparedness indices were used instead of a single factor. A very small number of indices could be identified to have been fully developed and with results freely available, namely the Rand Corporation's Infections Disease Vulnerability Index and the Global Burden of Disease's Healthcare and Quality Access Index, with more in the process of development (Global Health Security Index) or publication (Metabiota), at time of writing. Very little detail was available regarding methods or results from the two latter indices. Finally, the author was aware of the Joint External Evaluation (JEE), a WHO endorsed tool to determine country level preparedness. However, the number of countries taking part was too small to provide any insight into global preparedness (International Working Group on Financing Preparedness, 2017) and was therefore not used for this analysis.

The indices used for this specific analysis were therefore, those developed by Rand Corporation and the Global Burden of Disease. Each index was based on different factors and therefore represented health in different ways, as described in **Table 6.2**. Note that, as very limited information was available on the Global Health Security Index, this was not included in the **Table 6.2**. Additionally, the Rand Corporation index will thereafter be referred to as the 'Rand index' and the Healthcare Access and Quality Index as the 'GBD' index. To make both index results comparable, the GBD scores were divided by a factor of 100.

Although the preliminary results from both indices were very similar, their representation of healthcare systems were very different (**Table 6.2**), with the GBD index only considering diseases (both infectious and non) and the Rand index also considering governmental and demographic aspects of health. As it was unclear whether focusing on population health alone was a good manner of representing a country's healthcare system or not, it was therefore not possible to justify the use of one index over the other. Therefore, both (Rand and GBD) were used in turn for this analysis.

Table 6.1: Summary of the indicators (downloaded from the World Bank, data.worldbank.org), reasoning and drawbacks for each indicator used as a proxy for healthcare development.

Indicator name	Download date	Reasoning	Drawback
Incidence of tuberculosis (per 100,000 people)	21st February 2018	Provides an estimate of the burden of Tuberculosis infection in the population and the challenge for control programmes. Given the slow progress of the disease; control programmes are slower in reducing incidence than prevalence and mortality (World Health Organization; 2015b).	External funding is likely to have a large impact on certain countries; difficult to tell how much.
Antiretroviral therapy coverage (% of people living with HIV)	21st February 2018	Allows an understanding of the progress made to provide therapy to all infected (especially in low- and middle-income countries) (World Health Organization, 2015b).	External funding is likely to have a large impact on certain countries; difficult to tell how much.
Health expenditure, total (% of GDP)	21st February 2018	According to the WHO, in order for a country healthcare system to prevent infections, it must have a well-functioning system (World Health Organization, 2005).	External funding is likely to have a large impact on certain countries; difficult to tell how much. Difficult to dissociate governmental from out-of-pocket funding.
Immunization, measles (% of children ages 12-23 months)	6th March 2018	According to the WHO, immunization is crucial in reducing under-five mortality, and measles vaccination an indicator of healthcare system performance (World Health Organization, 2015b).	External funding is likely to have a large impact on certain countries; difficult to tell how much. Some developed countries (good healthcare systems) are seeing reduced uptakes.
Life expectancy at birth, total (years)	6th March 2018	This indicator captures both infectious and non-infectious causes of morbidity and mortality (World Health Organization, 2015b).	Does not only depend on healthcare system (for example, air pollution, war…)

Table 6.2: Brief summary of each preparedness index considered for this analysis and for which some information is known, including methods, factors considered, results and additional comments.
(For further information regarding the methods, it is advised to refer directly to the index description documents).

Source and index name	Aims	Methods (brief)	Factors considered	Results (3 most and least vulnerable)	Comments
Rand Corporation[1]– 'Infectious disease vulnerability index'	Provide a tool to help understand which countries are most vulnerable to a potential outbreak, not success or failure to respond.	Using a literature review search and professional experiences. Each indicator was weighted and summed into one index. Country rankings: 0 (worst) to 1 (best) Data years included: unclear Number of countries: 195	Demographics; health care; public health; disease dynamics; political-domestic and international; economic	Most vulnerable: Somalia, Central African Republic, Chad. Least vulnerable: Norway, Canada, Finland	Validates own weighting system and includes indirect influences on healthcare.
Global Burden of Disease [2] - 'Health Care Access and Quality Index'	To gain an understanding of global healthcare access and quality, and its changes between 1990 and 2015, from the Global Burden of Diseases, Injuries and Risk Factors Study 2015.	The group calculated "cause-specific mortality and population attributable fractions" taking into account a number of risk factors from Nolte and McKee. Country rankings: 0 (worst) to 100 (best) Data years included: 1990 to 2015 Number of countries: 195	Diarrhoeal diseases; respiratory infections (lower and upper, Tuberculosis); diphtheria; whooping cough; tetanus; measles; maternal disorders; neonatal disorders; cancer (non-melanoma skin; cervical; uterine; testicular; Hodgkin's lymphoma; leukaemia); heart disease (rheumatic; ischaemic; hypertensive); cerebrovascular disease; peptic ulcer disease; appendicitis; hernia; gallbladder and biliary diseases; epilepsy; diabetes mellitus; chronic kidney disease; congenital heart anomalies; adverse effect of medical treatment	Most vulnerable: Central African Republic, Afghanistan, Somalia. Least vulnerable: Andorra, Iceland, Switzerland.	Very detailed methods and results by country. Only focusses on disease (infectious and non) but not external factors. Reflects our previous finding that healthcare level varies according to indicator used.
Metabiota Preparedness Index[3]	Develop a measure of understanding of a country's ability to detect and respond to an outbreak.	Multidimensional framework based on index weighting then ranking . Country rankings: 5 (worst) to 1 (best) Data years included: unknown Number of countries: 188	Public health infrastructure; infrastructure (physical and communication); management capacities (bureaucratic and public); financial resources; risk communication	Most vulnerable: West and Central Africa, some of South East Asia. Least vulnerable: Western Europe, North America. Detailed results not yet available.	No more details on methods available (in process of publishing)

[1] Moore *et al.* (2016) and www.rand.org/pubs/research_reports/RR1605.html;
[2] Barber *et al.* (2017) and www.thelancet.com/journals/lancet/article/PIIS0140-6736(17)30818-8/fulltext;
[3] International Working Group on Financing Preparedness (2017) and www.metabiota.com/product.

Relative distance and regression analysis

To determine whether the two indices provided similar results in terms of healthcare access, the country scores according to each index were plotted against one another, before performing a Bland-Altman analysis (analysis of measuring agreement between two quantitative measurements; Giavarina (2015)). The latter method plots the mean of results for both methodologies (here the healthcare index scores per country) against the difference in results. Note that the GBD index scores were divided by a factor of 100 for this analysis only, to make the results more readily comparable.

To determine which countries might pose a higher risk for the global community, it was assumed that connectivity and the level of healthcare provided should be weighted evenly, and therefore the connectivity of each country was plotted against each healthcare index value. A point representing the 'worst case scenario' (WCS) showing the highest connectivity present in the data against a healthcare development index score of zero, was added to these plots. This point was, of course fictitious, but helped determine which countries were closest to such a scenario. The relative distance between each country and the WCS point was then calculated and represented using a radar plot, for each index. The equation used to calculate the relative distance of country *i* was:

$$relative\ distance_i = \sqrt{(x_i - 0)^2 + \ (1 - \frac{\theta_i}{\theta})^2}$$

Where θ is the fastest connectivity observed in all countries, θ_i is the connectivity for country *i* and x_i is the preparedness index value. In order to calculate these relative distance values, both axes were adjusted to be on comparable scales; ranging between 0 and 1 for both indices. The relative distance of each country from the WCS point was recorded in **Table 6.4,** along with each index score and connectivity.

To determine whether there was a relationship between each county's connectivity and healthcare index, a simple linear regression analysis was first attempted. However, when plotting residuals it became clear that the relationship was not linear, therefore a second order polynomial model was developed, with the equation:

$$x_i = a_0 + a_1\theta_i + \ a_2\theta_i^2 + \ \mathcal{E}$$

With $\mathrm{x_i}$ index score (GBD or Rand) for country *i*, θ_i is the connectivity for country *i*, with parameters *a* and Ɛ an error coefficient.

Results

The seasonality of the global connectivity (θ) can be seen in **Figure 6.1**, with some similar seasonal patterns to the global airline network seen in previous chapters clearly evident: peaks occurring between July and September and troughs between January and March, with additional peaks and troughs were seen in the intervening seasons. However, the average seasonal variations in global connectivity are relatively small, ranging from 0.03 (month 1) to 0.05 (month 55).

As can be seen in **Figure 6.1**, month 55 was the month with the quickest connectivity, so was used to determine the 2.5 and 97.5% percentiles for each country, as a means of representing the worst-case scenario of an outbreak spreading internationally, shown in **Figure 6.4**. Overall, the majority of countries showed consistent connectivity values, especially those with high index scores, with some variations. Additionally, some island countries like Fiji and Iceland, showed large variations in connectivity.

Overall, both healthcare indices showed similar results (correlation coefficient=0.90), with countries such as Somalia and the Central African Republic recorded as being some of the most vulnerable countries to the threat of an epidemic by both indices, whereas countries in Western Europe were most resilient (**Table 6.3**, **Figure 6.2**). The Rand index showed a large number of countries with healthcare scores of less than 0.50 (n=72, 38.5%) and the majority of these were in the Western/Central Africa region (n=24, 33%). The GBD index showed fewer countries with scores below 0.50 (n=47, 24%), with the majority of these (n=17, 36%) again in Western/Central Africa.

However, when considering the Bland-Altman analysis (**Figure 6.3**), most disagreement (largest differences between index scores for each country) was seen in the countries with the smallest index scores, such as Somalia and Central African Republic. Most agreement (smallest differences) was seen in countries with high index scores, represented closer to the red line of no difference. The largest difference between index scores was seen in Mauritania (difference= -0.41).

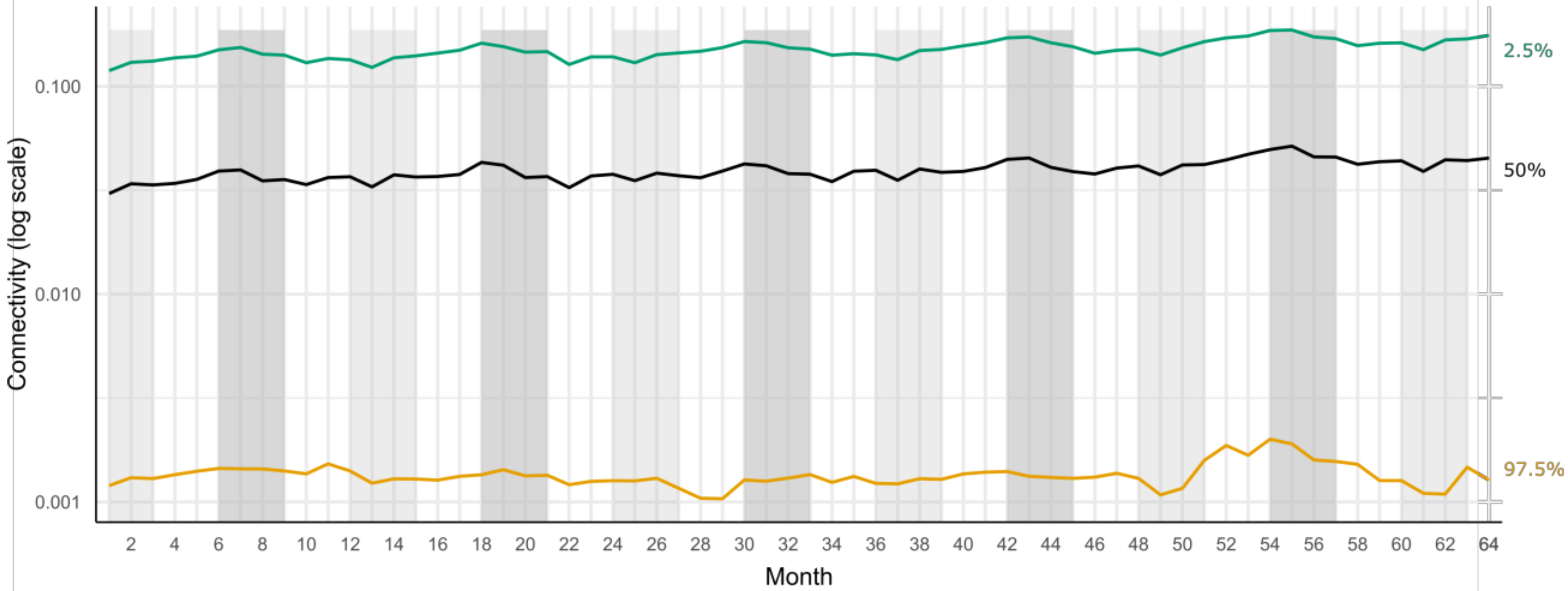


Figure 6.1: Global connectivity, mean, 2.5% and 97.5% percentiles (log scale), from the percolation model in monthly intervals, with Juy to September shadded in dark and January to March shadded in light.

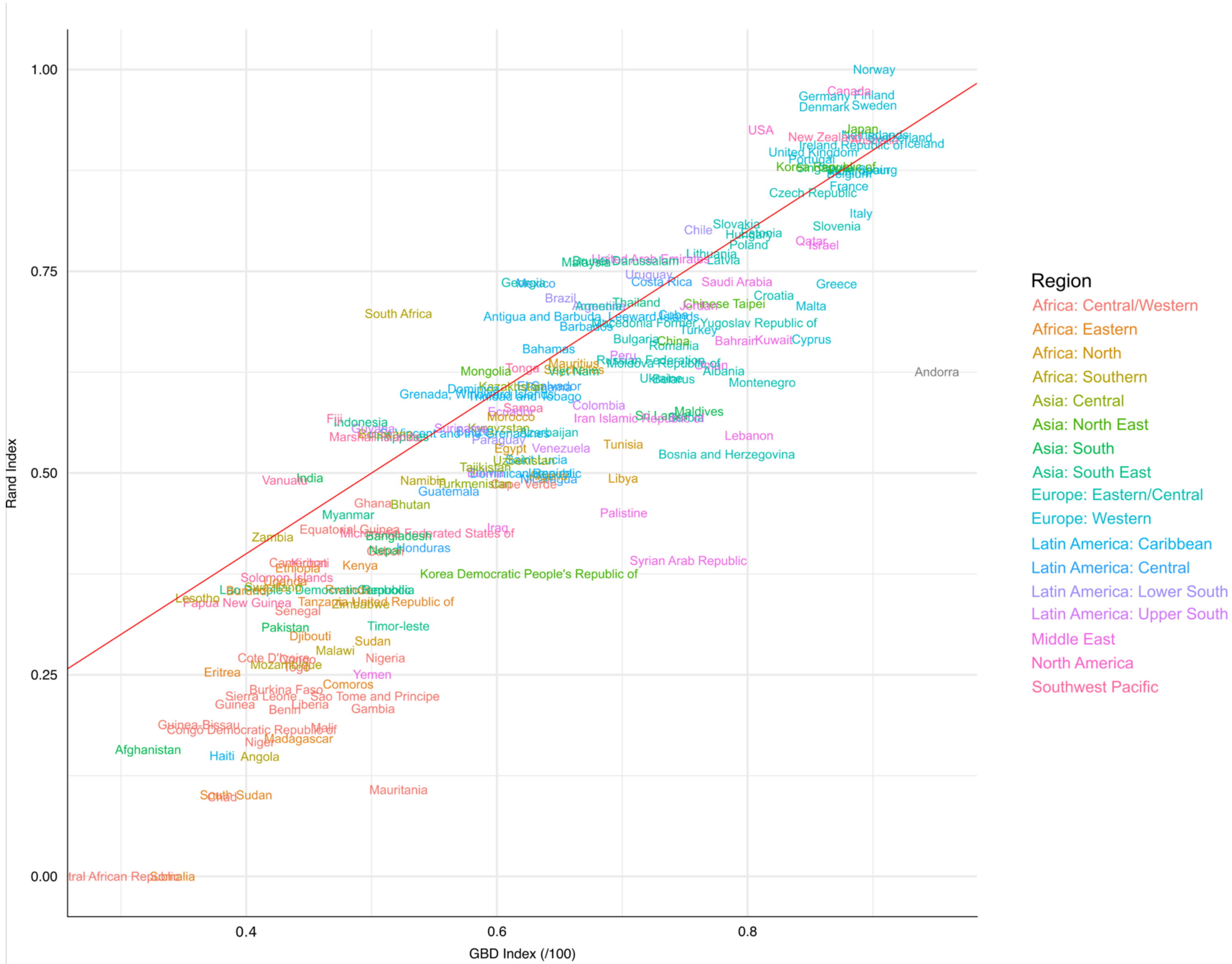


Figure 6.2: Direct comparison of prepardness indexes: Health Care and Quality Access (GBD) and Infectious Disease Vulnerability (Rand). Note: the GBD index scores have been divided by a factor of 100 to make the indexes comparable. The red line indicates the line of equality (x=y).

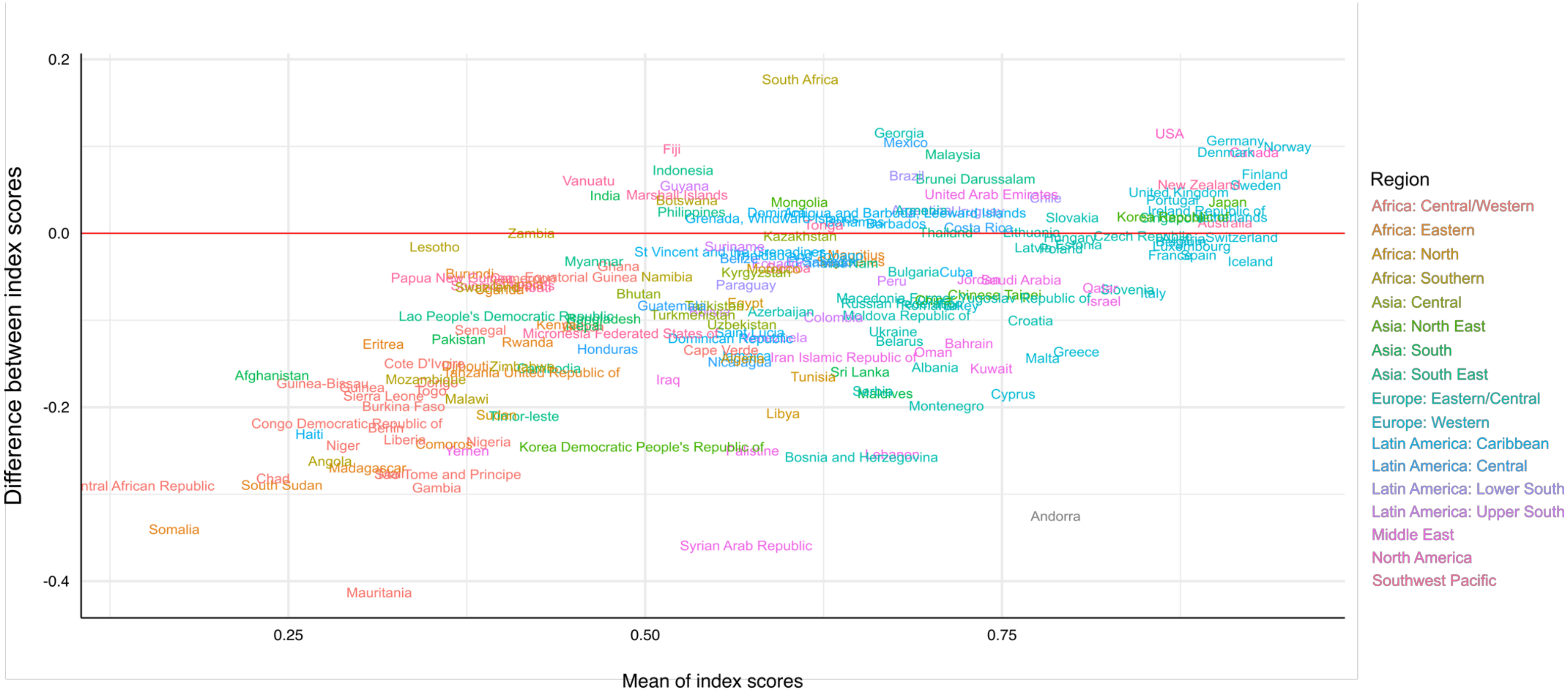
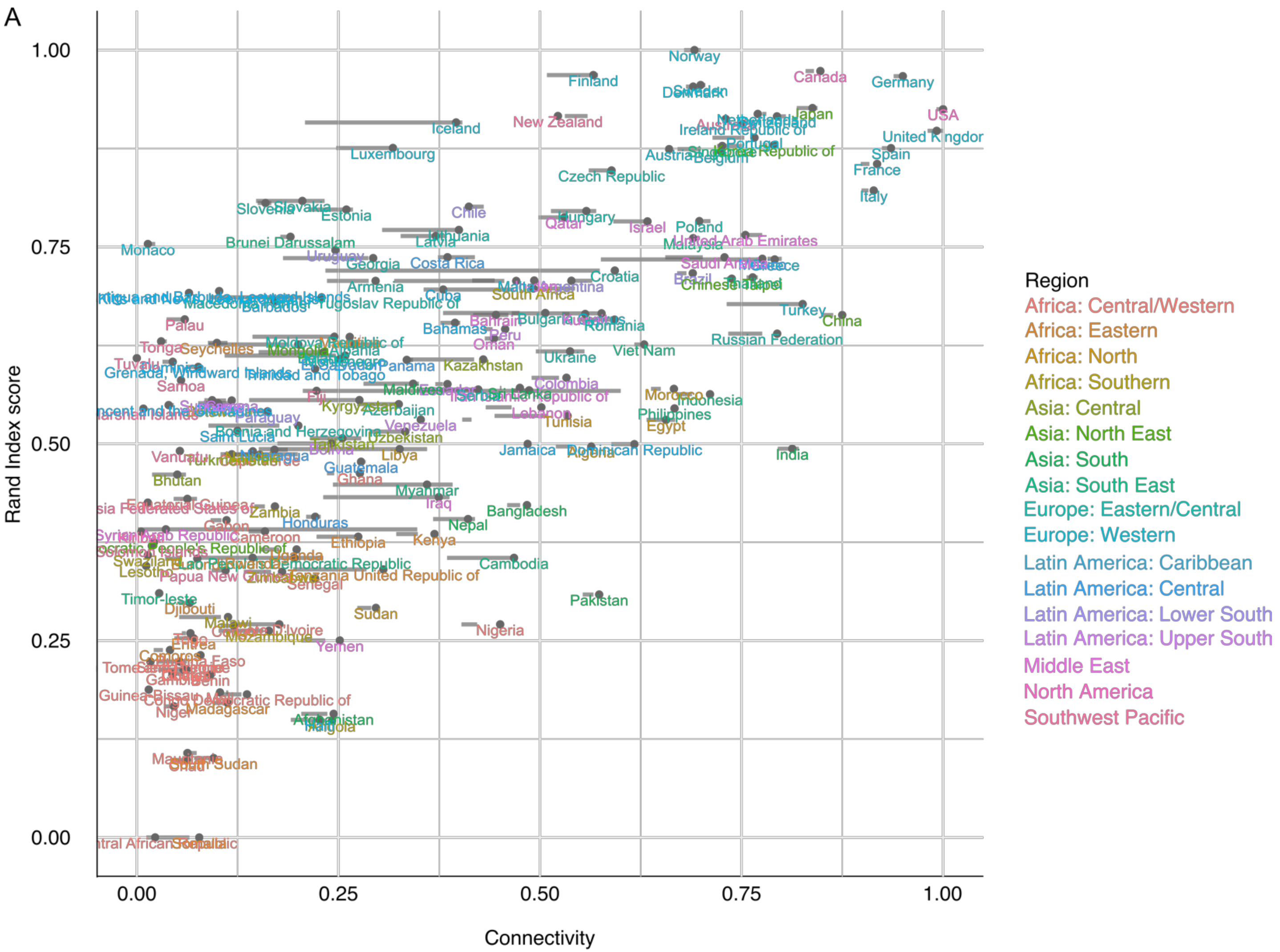


Figure 6.3: Bland Altman comparison plot of preparedness indexes.
Note: the GBD index scores have been divided by a factor of 100; the horizontal line represents the average difference between both indexes.

(Page intentionally left blank)



(Figure 6.4 continues on next page)

(Figure 6.4 continued)

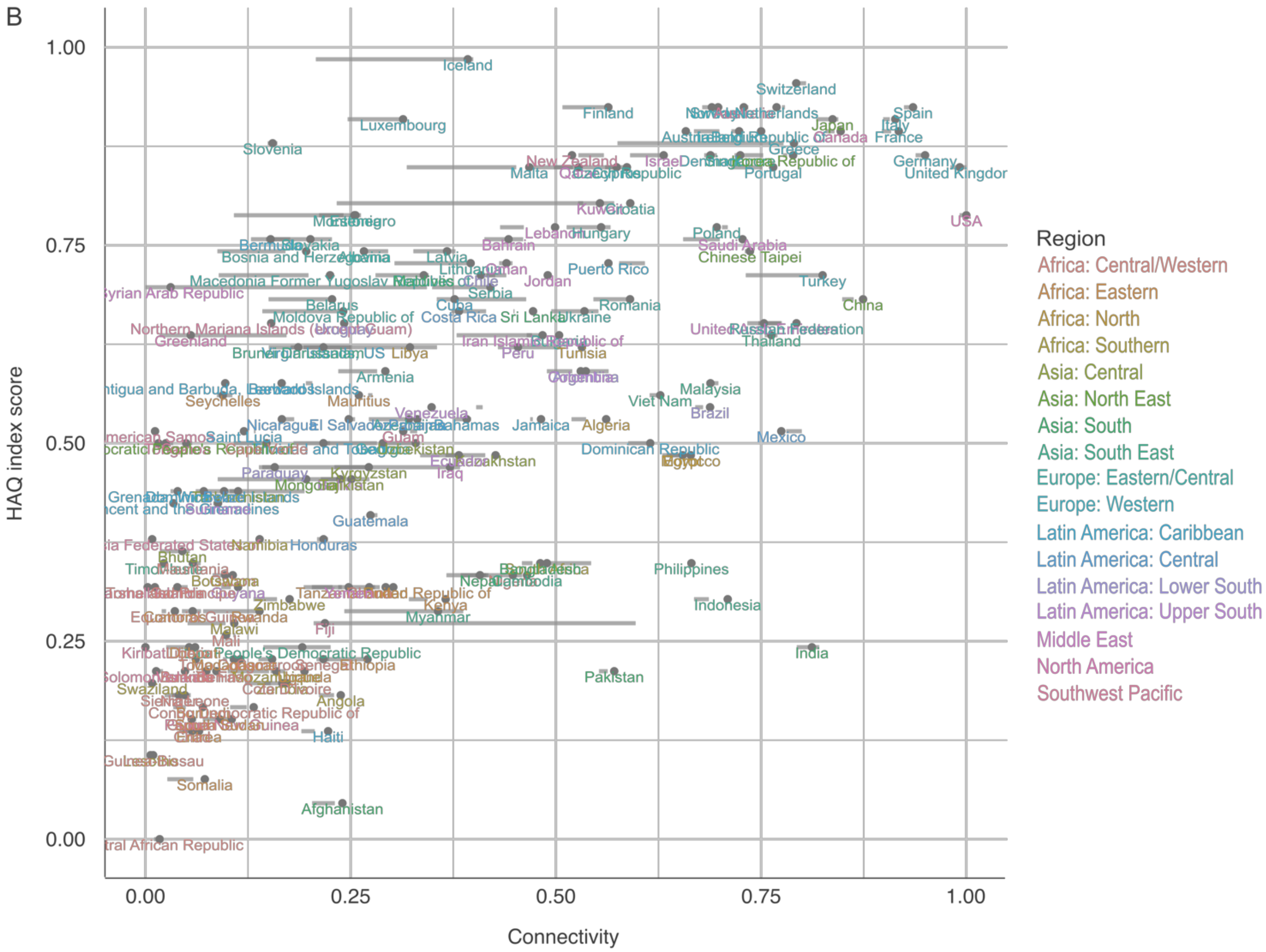


Figure 6.4: Country connectivity value (2.5% and 97.5% quantiles) for the month with the fastest global connectivity (month 55 from Figure 6.1) against each index score values (Rand index A and GBD index B), with each country grouped according to OAG region.
Note: The errorbars for each index score value could not be determined because of the nature of the data. The HAQ index score have been divided by a factor of 100.

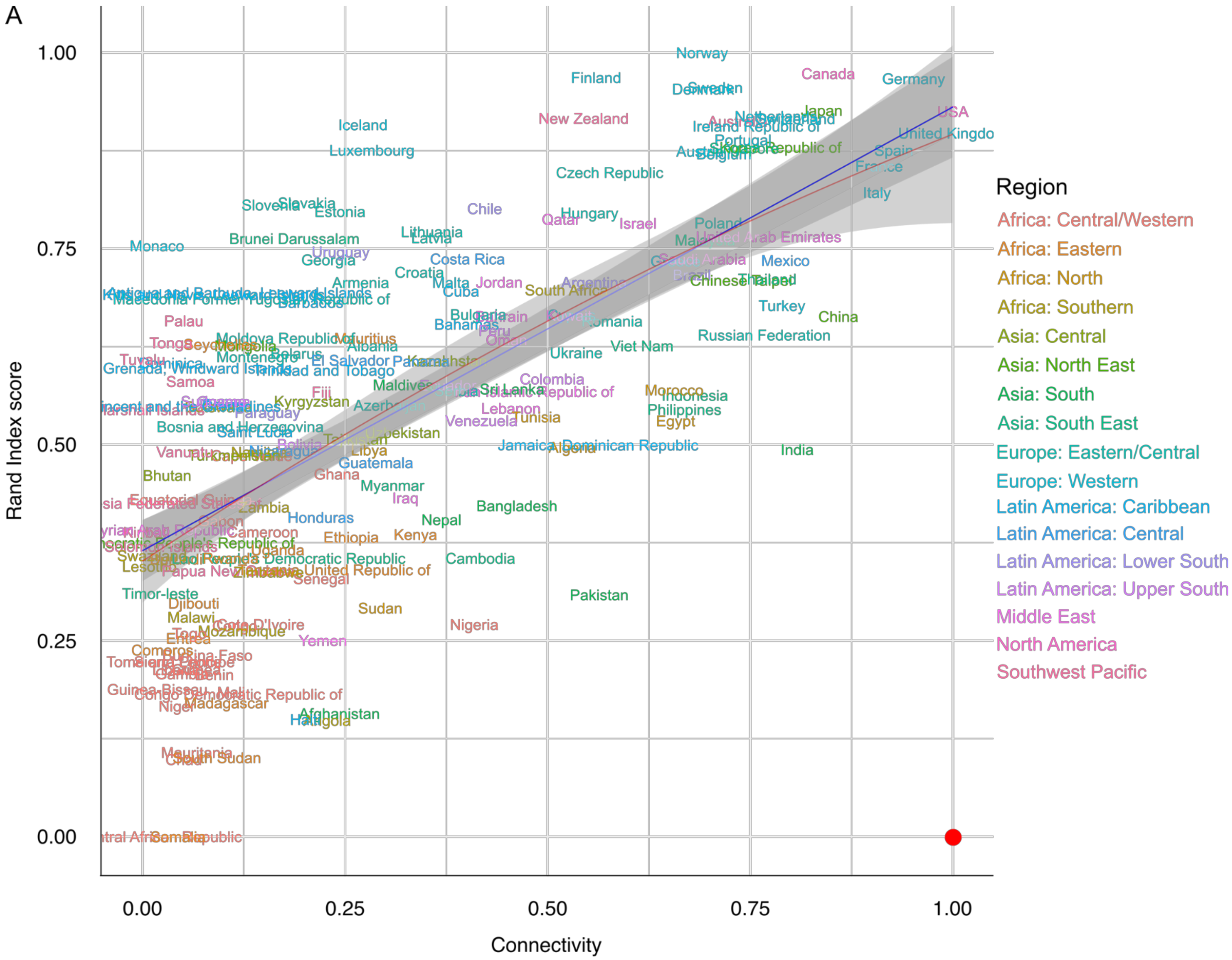
Figure 6.5 A showed the range of index scores for each country, from 0 (Somalia) to 1 (Norway). The USA was the country with the fastest connectivity (0.17 countries per time-step), whereas Tuvalu was the slowest (0.00071 countries per time-step). From **Figure 6.5 A**, it could be seen that countries with the lowest index scores also tended to be poorly connected, and that the majority of these were in Africa. Additionally, a number of small island nations (such as Tuvalu, Tonga and Samoa) had an average-to-good healthcare score but poor connectivity. A number of European countries such as Iceland and Luxembourg had good healthcare index scores but low connectivity scores (index=0.98, connectivity=0.047 and index=0.88, connectivity=0.049, respectively). Overall, Latin American countries (excluding Caribbean) and Middle Eastern countries had good healthcare and average connectivity, with some exceptions such as Honduras, which had a low healthcare score relative to other neighbouring countries (**Table 6.5**). The smaller islands of the Caribbean were not very well connected but had over average healthcare scores, except for Haiti, with poor healthcare and poor connectivity. On the other hand, Jamaica and especially the Dominican Republic had average healthcare and good connectivity. A group of six countries (France, Germany, Italy, Spain, United Kingdom and USA) were seen as having excellent connectivity and excellent healthcare. Countries such as India, Morocco, Egypt, Indonesia and the Dominican Republic were seen to have good connectivity but average healthcare scores, whereas Pakistan was seen to have a lower healthcare score than the latter countries.

Regarding the Global Burden of Disease's Healthcare Access and Quality Index, the smallest value was recorded in Central African Republic (29), and the highest in Andorra (95) although the latter was not represented in **Figure 6.6A** as this principality was not registered in the OAG data. Similar patterns and groupings of countries could be seen for both index scores (**Figure 6.5 A** and **Figure 6.6 A**), with African countries grouped together with low healthcare index scores and low connectivity, in direct opposition to Western European countries. Countries with good connectivity but average healthcare index scores or below were India, Philippines, Indonesia and Pakistan. Another grouping of countries could be seen with slightly above average healthcare and relatively good connectivity included the Dominican Republic, Viet Nam, Malaysia, Brazil, Egypt and Morocco.

From **Figures 6.5 B** and **6.6 B**, it became apparent that India was the closest to the Worst-Case Scenario point, thereby potentially posing the greatest risk to the global community (according to this analysis), followed by either Pakistan (Rand Index) or Indonesia (GBD Index). When considering the regional variations (black stars), both indices agreed that South Asia posed the highest risk to the global community (distance to WCS= 0.71 (Rand) and 0.77

(GBD)) (**Table 6.4**). In contrast, South West Pacific posed the lowest risk, with distance to WCS values of 1.06 (Rand) and 1.05 (GBD). This results from both the lower than average connectivity and higher than average level of healthcare development of South West Pacific nations (**Table 6.3** and **6.4**).

From the second order polynomial model (red line in **Figures 6.5A** and **6.6A**), the associated Akaike Information Criterion (AIC) values were smaller for the linear model: -109.42 for the linear model and -113.95 for the polynomial model using the Rand index; -96.93 for the linear model and -106.43 for the polynomial model using the GBD index.



(Figure 6.5 continues on next page)

(Figure 6.5 continued)

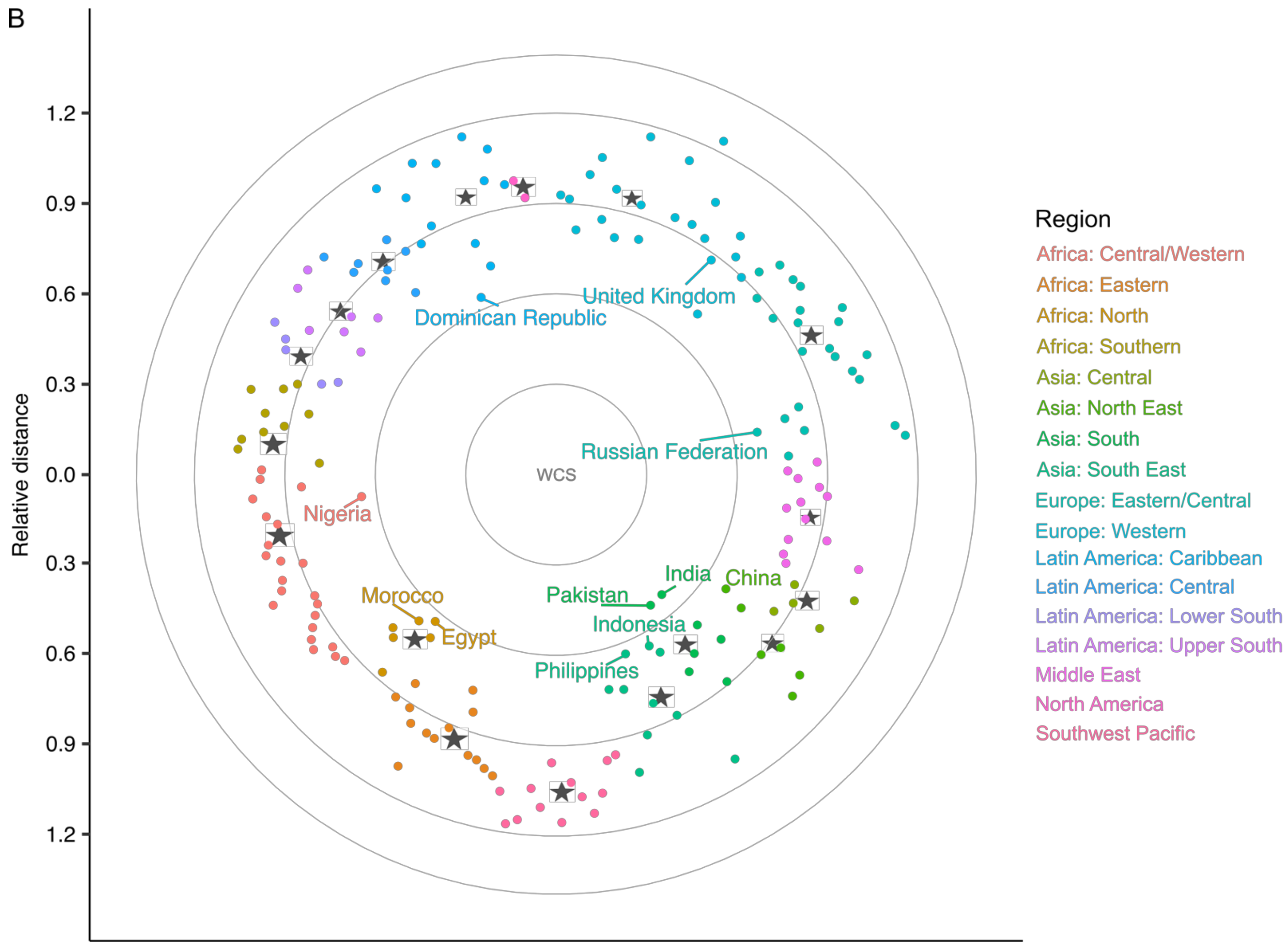
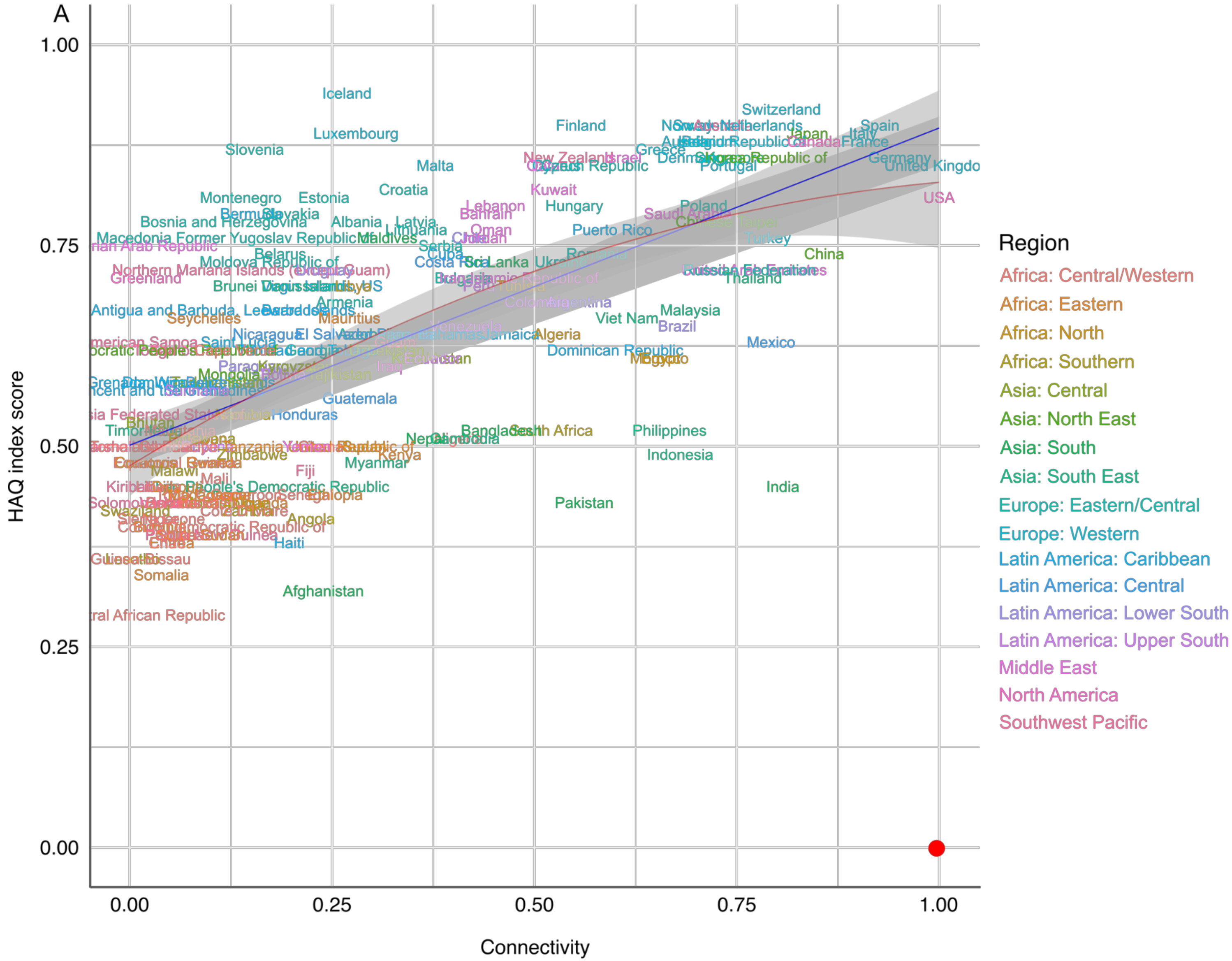


Figure 6.5: Using the Rand Index, (A) plot of country connectivity against index scores colour coded by OAG region, with linear regression line (blue) and second order binomial line (red); (B) radar plot of the relative distance of each country (grouped by OAG region) from the 'Worst Case Scenario' (WCS) in incremental steps with regional averages in black stars.
Note: the ten countries with the smallest relative distance (rank one to ten) have been named on (B), as well as the UK.



(Figure 6.6 continues on next page)

(Figure 6.6 continued)

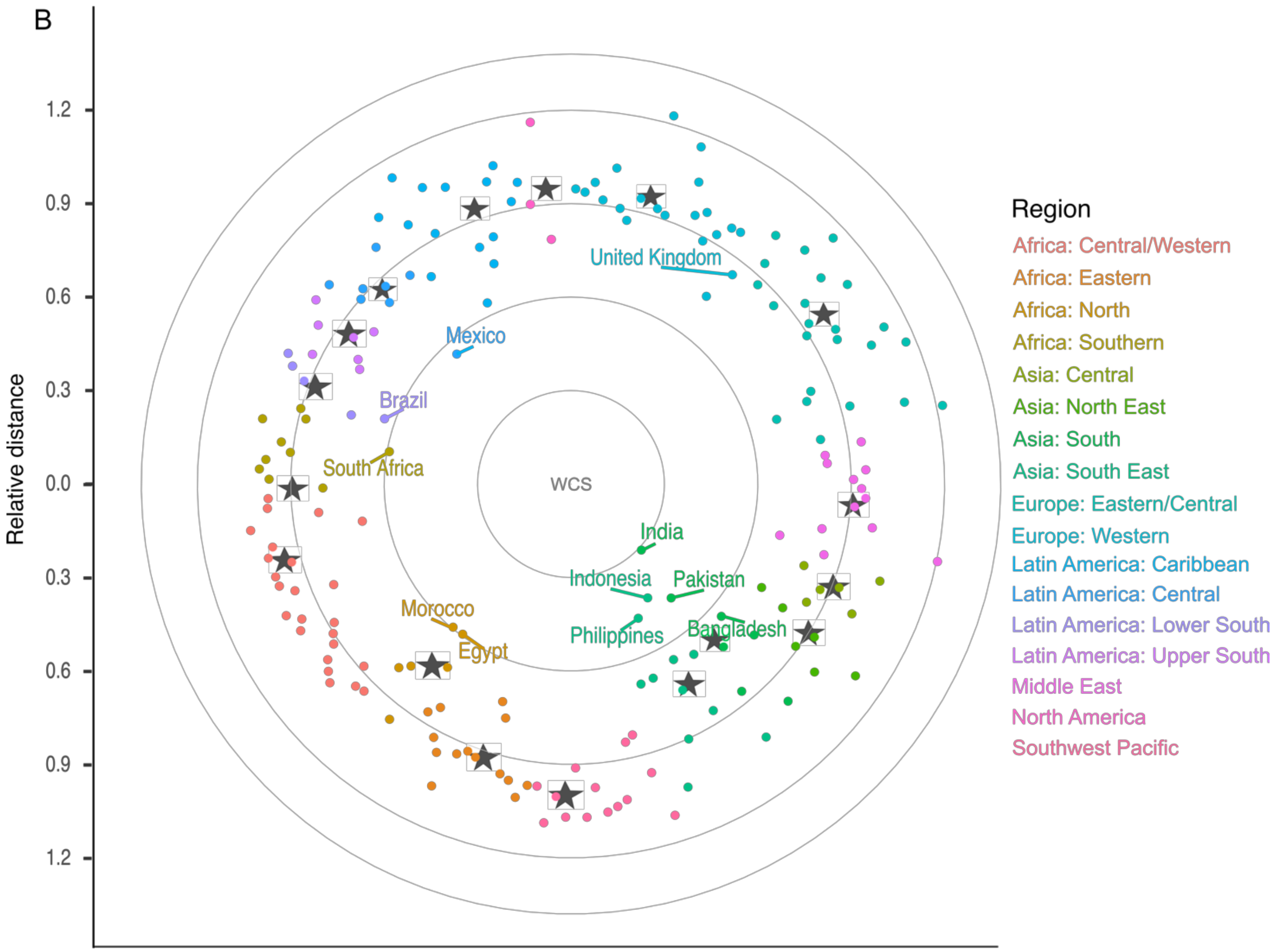


Figure 6.6: Using the Global Burden of Disease HAQ Index, (A) plot of country connectivity against HAQ index score colour coded by OAG region, with linear regression line (blue) and second order binomial line (red); (B) radar plot of the relative distance of each country (grouped by OAG region) from 'Worst Case Scenario' (WCS) in incremental steps with regional averages in black stars. Note: the ten countries with the smallest relative distance (rank one to ten) have been named (B), as well as the UK. HAQ score values have been divided by a factor of 100.

Table 6.3: Summary of the OAG region mean index scores, speed of connectivity (in countries per time step) and relative distance to WCS, according to index, ranked according to Rand ranking values.

Region name	GBD score	Rand score	Distance (GBD)	Distance (Rand)	Ranking (GBD)	Ranking (Rand)
South Asia	0.366	0.419	0.681	0.707	1	1
North Africa	0.510	0.486	0.740	0.727	2	2
South East Asia	0.466	0.574	0.744	0.822	3	3
Middle East	0.693	0.610	0.906	0.848	9	4
North Asia	0.692	0.697	0.898	0.900	7	5
Latin America: Upper South	0.491	0.564	0.872	0.905	6	6
Latin America: Central	0.500	0.577	0.863	0.915	5	7
Latin America: Lower South	0.594	0.702	0.853	0.935	4	8
East Africa	0.268	0.322	0.920	0.939	11	9
West/Central Africa	0.238	0.251	0.940	0.941	13	10
Central Asia	0.452	0.522	0.911	0.945	10	11
Western Europe	0.889	0.863	0.955	0.952	15	12
Southern Africa	0.253	0.389	0.898	0.953	8	13
North America	0.773	0.949	0.954	0.955	14	14
Eastern/Central Europe	0.712	0.693	0.980	0.965	16	15
Latin America: Caribbean	0.530	0.571	0.933	0.973	12	16
South West Pacific	0.447	0.572	0.999	1.066	17	17

Discussion

The aim of this study was to explore variations in the risk of international spread of an epidemic based on a composite score incorporating the given country's global connectivity and level of healthcare development. The results suggest that certain countries such as India and Pakistan, are well connected but with low-to-average healthcare systems, and as such may have high potential risk for international spread should an outbreak start in these countries. On the other hand, countries like Monaco and Slovenia were more likely to detect an outbreak early due to their good healthcare systems, but also had low connectivity, thereby better able to control its international spread.

The proxies initially considered to represent healthcare systems globally were all included in the World Health Organization's 2015 Global Reference List of 100 Core Health Indicators (World Health Organization, 2015a), and it was first assumed that these proxies were closely connected to a country's outbreak preparedness. However, each indicator had its flaws, such as the difficulty in determining the amount of foreign aid given for measles vaccination campaigns in resource poor countries (Gavi, 2017) from governmental vaccination campaigns, for example. Additionally, a country's spending on healthcare as a percentage of its GDP may also include external funding from the private sector, which may or may not be recorded and reported, depending on the country (World Health Organization, 2015b). The use of a single indicator to represent a healthcare system was determined not to be representative enough of a healthcare system for our purposes, and as has been suggested in previous studies (Chan *et al.*, 2013; Moore *et al.*, 2016), multi-factorial indices were used there-after. The two indices used here provided freely available results describing international healthcare systems, using factors such as population, health, demographics and economics among others (International Working Group on Financing Preparedness, 2017). Although this analysis was primarily focused on the international spread of infectious diseases, having an accurate representation of a country's healthcare system also included other factors such as demographics and economics. Therefore, it was difficult to determine which of the indices was the most appropriate to use.

It is interesting to note that countries which have seen important outbreaks develop into pandemics in this century alone were identified among those having good connectivity but also, interestingly, above average healthcare index scores. Of note were: Brazil from which the Zika pandemic spread across the American continent in 2015; China (including Hong

Kong) from which the Severe Acute Respiratory Syndrome (SARS) outbreak spread globally in 2003, and Mexico from which the H1N1 Influenza A strain also spread globally in 2009. These examples give some confidence that our model is able to predict which countries pose an international risk; however, it is worrisome that other countries like India, Pakistan, the Philippines and Indonesia, are equally well connected but with less developed healthcare systems than the countries previously listed. It might be reasonably expected that if an outbreak were to spark in these countries, the pathogen may be undetected for an extended period of time and could readily spread internationally, with the potential to spark further outbreaks. Some regions of the world may have good healthcare systems but be poorly connected, such as South West Pacific islands, resulting in a reduced risk to the global community as a consequence. At the regional level, there was agreement between both indices that Southern Asian countries (such as India) pose a higher risk to the global community, whereas the South West Pacific countries pose the lowest risk. This is likely a direct result of the connectivity of these locations, as South Asia is better connected than South West Pacific.

Whether a pathogen will spread internationally after initial importation and at what speed depends not only on the country of origin, but also on the epidemiological features of the pathogen. For example, vector borne pathogens will only spread into their imported destinations if the correct vectors are present (Tatem and Hay, 2007; Wilson, 1995). Additionally, the pathogen must be able to spread with relative ease within and between susceptible populations. Airborne pathogens, and especially those that are transmissible before the symptomatic phase, are particularly prone to global spread (Amesh *et al.*, 2018; Wilson, 1995). Finally, it can be argued that the pathogen must reach a part of the population that is able to fly internationally and come into contact with those who are infectious.

Civil unrest is known to impact a country's infrastructure, including healthcare and transportation (Bonds *et al.*, 2018). When considering where the next pandemic is most likely to spread from, having an understanding of the stability of a given country (political or otherwise) helps understand whether this risk is likely to change. The impact of war was seen in countries like Syria where the global connectivity was significantly reduced over the time period examined. The West African Ebola virus outbreak of 2014 showed the devastating impact an outbreak can have on a healthcare system and reverse hard-fought progress in a country's development. In 2014, it was reported by countries to WHO that only one-third of countries globally were suitably prepared to detect and respond to a national public health emergency, with African countries being the least prepared (International Working Group on

Financing Preparedness, 2017). A good and trusted healthcare system should be able to detect and control a pathogen in a timely manner, but will also disseminate correct and appropriate information to the population, thereby reducing the level of fear associated with an outbreak that may lead to large population movements (International Working Group on Financing Preparedness, 2017). In our strongly connected world, an outbreak in one country has the potential to cause public health concern internationally (Brower, 2003). Therefore, understanding global connectivity and preparedness is crucial.

Given the relatively large number of countries with poor-to-average healthcare but average-to-good connectivity, such as India and Pakistan, improving surveillance systems to quickly detect an outbreak is crucial to help avoid future pandemics. These countries must be encouraged (and/or supported by the international community) to further develop their healthcare system and preparedness to cope with the possibility of an outbreak without generating an international public health concern (International Working Group on Financing Preparedness, 2017). In the event of a public health event of international concern, the global community must unite to help control the spread of the given pathogen, however, this emergency help may be detrimental to the future development of a country by reducing the system's long term effectiveness and resilience (Harvard Global Health Institute, 2018). Therefore, international aid must be given in a manner that will encourage the development of a healthcare system, with staff training and infrastructure development (Bonds *et al.*, 2018).

In May 2018, the WHO and World Bank co-created the Global Preparedness Monitoring Board in the hope of improving global preparedness for the next pandemic, by holding all actors accountable for the development and maintenance of adequate healthcare systems. Additionally, the Board is placed in an ideal position to keep global health at the top of the international agenda, rather than letting other international issues take precedence and therefore continuing the pattern of responding only during international emergencies (Harvard Global Health Institute, 2018; World health Organization, 2018e). It could be argued that this analysis may help provide the new Board with the initial information regarding which countries should receive international funding as a priority.

Pandemics are known to be very costly. For example, the SARS outbreak of 2003 is estimated to have cost $52.2 billion to the global economy (International Working Group on Financing Preparedness, 2017). The most conservative models of future pandemic estimate the economic losses to be between 0.1 and 1.0% of the global GDP, which is relative to climate change and natural disasters (International Working Group on Financing Preparedness, 2017). Countries such as the United Kingdom regard the threat of a future pandemic as very serious, with the potential of causing thousands of extra deaths and important economic losses (United Kingdom Government, 2017). Given uncertainty about the timing of the next pandemic, global preparedness is all the more essential to reduce potential costs (International Working Group on Financing Preparedness, 2017; Semenza *et al.*, 2016b).

There is an abundance of literature stating that weak healthcare systems are ideal settings for an outbreak to spark due to slow within-country detection (Barber *et al.*, 2017; Bonds *et al.*, 2018; Elmahdawy *et al.*, 2017; Moore *et al.*, 2016). However, how these countries are connected globally and the potential international spread of an epidemic (thus developing into a pandemic) is not often considered together. Although developing healthcare systems globally is of crucial importance (and strongly encouraged by the Sustainable Development Goals (United Nations, 2017)), it can be argued that considering each country's pandemic spread potential must also be considered to prioritise and take early and adequate control measures and thereby prevent the further costs associated with a pandemic, both in terms of mortality, morbidity and economics. Given today's global connectivity and the unpredictability of the location of the start of an outbreak, ensuring a level of preparedness at a global level is critical (International Working Group on Financing Preparedness, 2017).

Although significant progress had been made globally in terms of access to and overall quality of healthcare provided by each country, the gap between countries with good and poor healthcare systems is widening. Some countries like Turkey, China and South Korea have improved their healthcare systems in a short period of time and now have good systems. Others such as Ethiopia, Peru and the Maldives have out-performed their expected rate of improvement (Barber *et al.*, 2017).

It is in the interest for the global community to invest in strong healthcare systems to prevent outbreaks from developing into pandemics, as these have a direct impact on economic growth (as seen in South Korea, and West Africa) and may overturn developmental progress already achieved (International Working Group on Financing Preparedness, 2017). Because a country is deemed vulnerable, that does not entail that an outbreak will start in that country or that it will fail in controlling it. But rather that control measures must be taken promptly, in a culturally sensitive manner and be adequately targeted given the pathogen and population (Moore *et al.*, 2016).

Limitations

A number of limitations were present in this analysis, including some relating to the index data available. Firstly, although a remarkable effort was made by the different groups to represent healthcare systems in the most accurate manner possible, there is always the possibility of bias and error that the author is not aware of, and that cannot be accounted for. Therefore, adding confidence intervals was not possible. Of the many factors included in each index, not many overlapped and which ones represent healthcare in a more accurate manner was unclear at time of writing. Even though there was a large overlap in the number of countries present in all data sets (global connectivity and both indices), a number were notably absent from various ones. The majority of those were small nations, such as Andorra and San Marino (missing from the airline data); however, the majority if not all of those missing from the indices were island nations, such as Aruba and Saint Kitts (missing from both indices). The absence of these nations may prove to be of importance if, like the Dominican Republic, they are well connected to the rest of the world but have a limited healthcare system. Additionally, these indices were done at the country level and did not consider within-country regions, which may have a significant impact on the results, especially for large countries like India. Only the Rand index included demographics in its factors, which is likely to also influence the location of the emergence of the next outbreak. Neither index used any measure of ecological changes within the country, such as level of deforestation, which have an important impact on the number of spill over events, from which outbreaks of zoonotic origin may spark (Jones *et al.*, 2008). Finally, the speed at which a pandemic would spread globally varies according to the pathogen in question and how contagious it is. For example, respiratory pathogens are difficult to control given their modes of transmission, and RNA viruses have high replication and mutation rates (Amesh *et al.*, 2018). Taking the pathogen type into account was partially considered by the GBD index, but not by the Rand index. Finally, this analysis only considered the airline spread of a pathogen, without considering other modes of transportation that cross international boundaries, such as sea and land travel. This could not be considered in the model as the only transportation data available to the author at the time was airline data.

Table 6.4: List of countries with OAG region, with scores assigned by each index and connectivity speed when used as seeding country and relative distance to 'WCS' according to index name, ordered by Rand ranking score.

Region name	Country name	GBD score	Rand score	Connectivity	Distance (GBD)	Distance (Rand)	Ranking (GBD)	Ranking (Rand)
Asia: South	India	0.45	0.49	0.812	0.307	0.530	1	1
Asia: South	Pakistan	0.43	0.31	0.571	0.478	0.534	3	2
Africa: North	Egypt	0.61	0.53	0.654	0.595	0.631	7	3
Asia: South East	Philippines	0.52	0.55	0.665	0.483	0.638	4	4
Latin America: Caribbean	Dominican Republic	0.62	0.50	0.615	0.631	0.639	12	5
Asia: South East	Indonesia	0.49	0.56	0.709	0.420	0.647	2	6
Africa: Central/Western	Nigeria	0.51	0.27	0.448	0.645	0.650	13	7
Africa: North	Morocco	0.61	0.57	0.664	0.590	0.666	6	8
Asia: North East	China	0.74	0.66	0.874	0.693	0.678	20	9
Europe: Eastern/Central	Russian Federation	0.72	0.64	0.793	0.684	0.681	18	10
Asia: South East	Cambodia	0.51	0.36	0.465	0.630	0.683	11	11
Africa: North	Algeria	0.64	0.50	0.561	0.688	0.684	19	12
Asia: South	Bangladesh	0.52	0.42	0.481	0.625	0.684	9	13
Europe: Western	Turkey	0.76	0.68	0.825	0.733	0.710	24	14
Latin America: Caribbean	Jamaica	0.64	0.50	0.482	0.741	0.726	25	15
Asia: South East	Viet Nam	0.66	0.63	0.627	0.673	0.734	15	16
Africa: North	Tunisia	0.7	0.54	0.532	0.778	0.742	33	17
Asia: South East	Thailand	0.71	0.71	0.763	0.679	0.747	16	18
Asia: South	Nepal	0.51	0.40	0.408	0.680	0.750	17	19
Asia: North East	Chinese Taipei	0.78	0.71	0.736	0.788	0.756	37	20
Latin America: Central	Mexico	0.63	0.74	0.775	0.562	0.763	5	21
Africa: North	Sudan	0.5	0.29	0.293	0.776	0.764	32	22
Latin America: Upper South	Colombia	0.68	0.58	0.530	0.755	0.765	28	23
Africa: Eastern	Kenya	0.49	0.39	0.366	0.703	0.767	22	24
Middle East	Iran Islamic Republic of	0.71	0.57	0.484	0.819	0.768	48	25
Middle East	Lebanon	0.8	0.55	0.499	0.921	0.772	96	26
Europe: Eastern/Central	Ukraine	0.73	0.62	0.535	0.813	0.773	45	27
Asia: South	Afghanistan	0.32	0.16	0.240	0.761	0.773	29	28

(Table 6.4 continues on next page)

(Table 6.4 continued)

Region name	Country name	GBD score	Rand score	Connectivity	Distance (GBD)	Distance (Rand)	Ranking (GBD)	Ranking (Rand)
Europe: Eastern/Central	Romania	0.74	0.66	0.590	0.795	0.781	41	29
Latin America: Lower South	Brazil	0.65	0.72	0.688	0.628	0.786	10	30
Africa: Southern	Angola	0.41	0.15	0.238	0.783	0.786	34	31
Latin America: Upper South	Venezuela	0.65	0.53	0.349	0.850	0.787	58	32
Asia: South	Sri Lanka	0.73	0.57	0.472	0.850	0.789	59	33
Middle East	United Arab Emirates	0.72	0.77	0.753	0.697	0.798	21	34
Middle East	Saudi Arabia	0.79	0.74	0.727	0.805	0.799	42	35
Middle East	Iraq	0.6	0.43	0.371	0.785	0.802	35	36
Europe: Western	Greece	0.87	0.73	0.790	0.904	0.810	81	37
Latin America: Caribbean	Haiti	0.38	0.15	0.223	0.789	0.813	39	38
Europe: Western	Cyprus	0.85	0.67	0.574	0.949	0.815	120	39
Middle East	Kuwait	0.82	0.67	0.554	0.919	0.816	92	40
Middle East	Yemen	0.5	0.25	0.248	0.817	0.817	47	41
Asia: South East	Malaysia	0.67	0.76	0.688	0.655	0.820	14	42
Asia: South East	Myanmar	0.48	0.45	0.356	0.705	0.824	23	43
Europe: Western	Italy	0.89	0.82	0.913	0.913	0.827	89	44
Latin America: Lower South	Argentina	0.68	0.71	0.536	0.751	0.834	27	45
Europe: Eastern/Central	Poland	0.8	0.78	0.696	0.830	0.835	51	46
Africa: Eastern	Ethiopia	0.44	0.38	0.271	0.764	0.835	30	47
Europe: Eastern/Central	Serbia	0.75	0.57	0.420	0.907	0.836	83	48
Africa: Eastern	Tanzania United Republic of	0.5	0.34	0.302	0.767	0.836	31	49
Middle East	Oman	0.77	0.63	0.440	0.918	0.840	90	50
Africa: Southern	South Africa	0.52	0.70	0.489	0.618	0.845	8	51
Africa: Central/Western	Senegal	0.44	0.33	0.217	0.815	0.846	46	52
Latin America: Upper South	Ecuador	0.61	0.58	0.382	0.786	0.848	36	53
Asia: Central	Uzbekistan	0.62	0.52	0.329	0.837	0.853	54	54
Latin America: Central	Guatemala	0.56	0.48	0.274	0.833	0.857	52	55
Latin America: Upper South	Peru	0.7	0.65	0.454	0.827	0.858	49	56
Europe: Western	France	0.88	0.86	0.918	0.898	0.860	79	57
Middle East	Bahrain	0.79	0.66	0.442	0.941	0.866	114	58

(Table 6.4 continues on next page)

(Table 6.4 continued)

Region name	Country name	GBD score	Rand score	Connectivity	Distance (GBD)	Distance (Rand)	Ranking (GBD)	Ranking (Rand)
Asia: Central	Kazakhstan	0.61	0.61	0.427	0.751	0.871	26	59
Africa: North	Libya	0.7	0.49	0.322	0.919	0.873	93	60
Middle East	Israel	0.86	0.78	0.631	0.939	0.874	112	61
Europe: Western	Spain	0.9	0.88	0.935	0.927	0.878	103	62
Latin America: Central	Honduras	0.54	0.41	0.217	0.870	0.880	69	63
Europe: Eastern/Central	Azerbaijan	0.64	0.55	0.321	0.861	0.886	64	64
Europe: Eastern/Central	Bulgaria	0.71	0.67	0.504	0.807	0.887	43	65
Latin America: Caribbean	Bahamas	0.64	0.65	0.391	0.807	0.887	44	66
Africa: Central/Western	Ghana	0.5	0.46	0.273	0.794	0.890	40	67
Asia: South	Maldives	0.76	0.58	0.339	0.972	0.890	134	68
Asia: South East	Lao People's Democratic Republic	0.45	0.36	0.191	0.844	0.893	55	69
Latin America: Central	Panama	0.64	0.61	0.331	0.854	0.894	61	70
Asia: Central	Tajikistan	0.59	0.51	0.251	0.876	0.895	72	71
Africa: Central/Western	Cote D'Ivoire	0.42	0.27	0.172	0.851	0.896	60	72
Europe: Western	United Kingdom	0.85	0.90	0.992	0.849	0.897	57	73
Africa: Central/Western	Congo Democratic Republic of	0.4	0.18	0.132	0.884	0.901	75	74
Middle East	Jordan	0.76	0.71	0.490	0.876	0.902	71	75
Asia: North East	Korea Republic of	0.86	0.88	0.789	0.889	0.905	77	76
Africa: Southern	Zimbabwe	0.49	0.34	0.176	0.878	0.909	73	77
Africa: Central/Western	Mali	0.46	0.18	0.098	0.938	0.909	110	78
Africa: Eastern	Uganda	0.43	0.37	0.194	0.834	0.910	53	79
Africa: Eastern	Madagascar	0.44	0.17	0.108	0.921	0.913	95	80
Europe: Eastern/Central	Hungary	0.8	0.80	0.555	0.892	0.913	78	81
Africa: Eastern	South Sudan	0.39	0.10	0.090	0.922	0.914	98	82
Asia: South East	Singapore	0.86	0.88	0.725	0.906	0.915	82	83
Europe: Western	Belgium	0.88	0.87	0.723	0.936	0.916	109	84
Africa: Southern	Mozambique	0.43	0.26	0.159	0.867	0.916	67	85
Latin America: Caribbean	Cuba	0.74	0.70	0.377	0.924	0.923	101	86
Middle East	Qatar	0.85	0.79	0.528	0.971	0.924	133	87

(Table 6.4 continues on next page)

(Table 6.4 continued)

Region name	Country name	GBD score	Rand score	Connectivity	Distance (GBD)	Distance (Rand)	Ranking (GBD)	Ranking (Rand)
North America	USA	0.81	0.93	1.000	0.788	0.925	38	88
Europe: Western	Portugal	0.85	0.89	0.765	0.880	0.926	74	89
Africa: Central/Western	Congo	0.44	0.27	0.116	0.913	0.926	88	90
Europe: Western	Austria	0.88	0.87	0.658	0.957	0.928	126	91
Africa: Central/Western	Benin	0.43	0.21	0.086	0.938	0.934	111	92
Europe: Western	Switzerland	0.92	0.92	0.793	0.977	0.936	140	93
Africa: Central/Western	Cameroon	0.44	0.39	0.154	0.876	0.936	70	94
Europe: Western	Ireland Republic of	0.88	0.91	0.750	0.928	0.938	104	95
Africa: Central/Western	Mauritania	0.52	0.11	0.058	1.005	0.939	153	96
Europe: Western	Malta	0.85	0.71	0.468	1.002	0.940	152	97
Asia: North East	Japan	0.89	0.93	0.837	0.924	0.940	99	98
Europe: Western	Netherlands	0.9	0.92	0.769	0.953	0.945	122	99
Europe: Eastern/Central	Czech Republic	0.85	0.85	0.587	0.944	0.947	116	100
Africa: Southern	Zambia	0.42	0.42	0.167	0.856	0.948	62	101
Africa: Central/Western	Burkina Faso	0.43	0.23	0.074	0.950	0.949	121	102
Latin America: Upper South	Bolivia	0.59	0.50	0.237	0.888	0.949	76	103
Latin America: Central	Costa Rica	0.73	0.74	0.382	0.909	0.950	87	104
Southwest Pacific	Australia	0.9	0.91	0.729	0.963	0.950	128	105
Europe: Eastern/Central	Albania	0.78	0.63	0.266	1.044	0.951	171	106
Africa: Central/Western	Chad	0.38	0.10	0.057	0.953	0.954	123	107
Africa: Eastern	Somalia	0.34	0.00	0.072	0.931	0.956	107	108
Southwest Pacific	Papua New Guinea	0.39	0.34	0.105	0.908	0.956	85	109
Africa: Central/Western	Guinea	0.39	0.21	0.056	0.956	0.958	125	110
Africa: Eastern	Rwanda	0.48	0.36	0.139	0.908	0.959	86	111
Latin America: Central	El Salvador	0.64	0.61	0.248	0.920	0.960	94	112
Latin America: Central	Nicaragua	0.64	0.49	0.166	0.988	0.961	144	113
Southwest Pacific	Fiji	0.47	0.57	0.219	0.828	0.964	50	114
Africa: Eastern	Mauritius	0.66	0.64	0.260	0.928	0.964	105	115
Asia: Central	Kyrgyzstan	0.6	0.56	0.272	0.866	0.966	66	116
Europe: Western	Germany	0.86	0.97	0.950	0.865	0.968	65	117

(Table 6.4 continues on next page)

(Table 6.4 continued)

Region name	Country name	GBD score	Rand score	Connectivity	Distance (GBD)	Distance (Rand)	Ranking (GBD)	Ranking (Rand)
Africa: Central/Western	Niger	0.41	0.17	0.041	0.976	0.972	139	118
Africa: Central/Western	Gambia	0.5	0.21	0.039	1.012	0.974	160	119
Europe: Eastern/Central	Croatia	0.82	0.72	0.591	0.901	0.976	80	120
Africa: Eastern	Eritrea	0.38	0.25	0.065	0.945	0.976	117	121
Africa: Central/Western	Togo	0.44	0.26	0.061	0.966	0.977	130	122
Latin America: Caribbean	Trinidad and Tobago	0.62	0.60	0.217	0.929	0.978	106	123
Africa: Central/Western	Central African Republic	0.29	0.00	0.017	0.983	0.979	142	124
Africa: Southern	Malawi	0.47	0.28	0.108	0.933	0.981	108	125
Africa: Central/Western	Sierra Leone	0.41	0.22	0.048	0.969	0.982	131	126
Africa: Central/Western	Liberia	0.45	0.21	0.053	0.977	0.983	141	127
Africa: Eastern	Djibouti	0.45	0.30	0.061	0.970	0.983	132	128
North America	Canada	0.88	0.97	0.847	0.907	0.985	84	129
Africa: Southern	Namibia	0.54	0.49	0.139	0.941	0.987	113	130
Latin America: Lower South	Chile	0.76	0.80	0.409	0.926	0.988	102	131
Africa: Central/Western	Gabon	0.51	0.40	0.106	0.954	0.989	124	132
Africa: Central/Western	Cape Verde	0.62	0.49	0.148	0.988	0.993	145	133
Europe: Eastern/Central	Latvia	0.78	0.76	0.367	0.975	0.999	138	134
Africa: Central/Western	Guinea-Bissau	0.36	0.19	0.009	0.996	0.999	148	135
Europe: Western	Sweden	0.9	0.96	0.698	0.972	1.000	135	136
Europe: Western	Denmark	0.86	0.95	0.688	0.918	1.002	91	137
Africa: Eastern	Comoros	0.48	0.24	0.036	1.006	1.004	156	138
Latin America: Lower South	Paraguay	0.6	0.54	0.157	0.965	1.004	129	139
Latin America: Caribbean	Saint Lucia	0.63	0.52	0.120	1.020	1.004	162	140
Europe: Eastern/Central	Lithuania	0.77	0.77	0.396	0.945	1.004	119	141
Africa: Central/Western	Sao Tome and Principe	0.5	0.22	0.012	1.038	1.010	170	142
Asia: Central	Turkmenistan	0.58	0.49	0.113	0.990	1.012	147	143
Europe: Eastern/Central	Armenia	0.68	0.71	0.292	0.922	1.017	97	144
Europe: Eastern/Central	Belarus	0.74	0.62	0.227	1.030	1.018	168	145
Africa: Eastern	Burundi	0.4	0.35	0.070	0.945	1.022	118	146
Europe: Eastern/Central	Bosnia and Herzegovina	0.78	0.52	0.196	1.094	1.023	183	147

(Table 6.4 continues on next page)

(Table 6.4 continued)

Region name	Country name	GBD score	Rand score	Connectivity	Distance (GBD)	Distance (Rand)	Ranking (GBD)	Ranking (Rand)
Southwest Pacific	New Zealand	0.86	0.92	0.520	0.988	1.023	146	148
Africa: Central/Western	Equatorial Guinea	0.48	0.43	0.058	0.985	1.034	143	149
Europe: Eastern/Central	Moldova Republic of	0.73	0.64	0.241	1.010	1.041	158	150
Southwest Pacific	Solomon Islands	0.43	0.37	0.013	1.009	1.045	157	151
Latin America: Caribbean	Barbados	0.67	0.68	0.166	1.013	1.045	161	152
Asia: North East	Korea Democratic People's Republic of	0.62	0.38	0.016	1.104	1.046	186	153
Europe: Western	Norway	0.9	1.00	0.690	0.975	1.047	136	154
Africa: Southern	Lesotho	0.36	0.35	0.006	0.999	1.050	150	155
Africa: Southern	Swaziland	0.42	0.36	0.008	1.011	1.051	159	156
Middle East	Syrian Arab Republic	0.75	0.39	0.031	1.194	1.051	192	157
Europe: Eastern/Central	Montenegro	0.81	0.61	0.254	1.085	1.054	182	158
Latin America: Central	Belize	0.58	0.55	0.096	1.005	1.056	154	159
Latin America: Upper South	Guyana	0.5	0.56	0.113	0.942	1.057	115	160
Africa: Southern	Botswana	0.51	0.55	0.097	0.962	1.059	127	161
Latin America: Lower South	Uruguay	0.72	0.75	0.242	0.999	1.062	149	162
Europe: Western	Finland	0.9	0.97	0.564	1.022	1.064	164	163
Europe: Eastern/Central	Georgia	0.62	0.74	0.289	0.869	1.065	68	164
Southwest Pacific	Vanuatu	0.43	0.49	0.048	0.975	1.067	137	165
Latin America: Upper South	Suriname	0.57	0.56	0.088	1.005	1.068	155	166
Southwest Pacific	Kiribati	0.45	0.39	0.000	1.029	1.068	167	167
Southwest Pacific	Micronesia Federated States of	0.54	0.43	0.008	1.062	1.073	176	168
Asia: Central	Bhutan	0.53	0.46	0.045	1.022	1.073	163	169
Asia: North East	Mongolia	0.59	0.63	0.196	0.924	1.074	100	170
Europe: Eastern/Central	Estonia	0.81	0.80	0.256	1.084	1.099	181	171
Africa: Eastern	Seychelles	0.66	0.63	0.094	1.065	1.101	177	172
Latin America: Caribbean	St Vincent and the Grenadines	0.57	0.55	0.034	1.055	1.104	174	173
Europe: Eastern/Central	Macedonia Former Yugoslav Republic of	0.76	0.69	0.225	1.053	1.105	173	174
Southwest Pacific	Samoa	0.62	0.58	0.050	1.074	1.105	180	175

(Table 6.4 continues on next page)

(Table 6.4 continued)

Region name	Country name	GBD score	Rand score	Connectivity	Distance (GBD)	Distance (Rand)	Ranking (GBD)	Ranking (Rand)
Latin America: Caribbean	Grenada, Windward Islands	0.58	0.60	0.071	1.028	1.107	166	176
Asia: South East	Brunei Darussalam	0.7	0.76	0.186	1.024	1.115	165	177
Latin America: Caribbean	Antigua and Barbuda, Leeward Islands	0.67	0.69	0.097	1.071	1.121	179	178
Southwest Pacific	Marshall Islands	0.5	0.55	0.003	1.047	1.131	172	179
Europe: Western	Luxembourg	0.89	0.88	0.314	1.139	1.132	189	180
Europe: Eastern/Central	Slovakia	0.79	0.81	0.201	1.101	1.135	185	181
Latin America: Caribbean	Dominica	0.58	0.60	0.039	1.056	1.138	175	182
Southwest Pacific	Tonga	0.62	0.63	0.025	1.096	1.152	184	183
Europe: Western	Iceland	0.94	0.91	0.393	1.157	1.164	191	184
Europe: Eastern/Central	Slovenia	0.87	0.81	0.155	1.219	1.165	193	185
Southwest Pacific	Guam	0.63	NA	0.314	0.858	NA	63	186
Southwest Pacific	American Samoa	0.63	NA	0.012	1.114	NA	187	187
Latin America: Caribbean	Virgin Islands, US	0.7	NA	0.217	1.000	NA	151	188
North America	Greenland	0.71	NA	0.055	1.139	NA	190	189
Southwest Pacific	Northern Mariana Islands (except Guam)	0.72	NA	0.153	1.068	NA	178	190
Latin America: Caribbean	Puerto Rico	0.77	NA	0.564	0.848	NA	56	191
Latin America: Caribbean	Bermuda	0.79	NA	0.153	1.137	NA	188	192
NA	Andorra	0.95	0.63	NA	NA	NA	194	193
NA	Palestine	0.7	NA	NA	NA	NA		194
NA	Timor-Leste	0.52	NA	NA	NA	NA		195

Chapter 7 – Discussion and conclusion

Preamble

This final chapter summaries key findings of the thesis as well as their implications in terms of research and suggests possible future work.

Summary of findings

In the 21st century alone, the global community has seen a number of outbreaks develop into pandemics, such as SARS (2003), Influenza A H1N1 (2009) or Zika virus (2016), each with an important cost, both in terms of lives lost (or impacted) and economics. With ever more airline passengers travelling today, modelling international travel is an important tool that is being increasingly used and referred to by policy makers. The aims of this thesis were to understand what the airline data represents in terms of global airline passenger movements and determine whether its use was appropriate to understand the international spread of human infectious diseases.

From the systematic review it became apparent that expensive and closed-source data sets such as IATA and OAG were most often used by researchers to model the international spread of human infectious diseases. These data sets are sold (sometimes at a very high cost) by the commercial airline industry as highly accurate airline data between international airports (OAG, 2013). Access to this global representation of passengers comes with a license for a given length of time and user restrictions but shows limited geographical and temporal restrictions. However, the financial cost of these data sets may be a barrier for some research groups, making open access data a more suitable alternative. Examples of such data providers include the US Department of Transport and the UK's Office for National Statistics. Although it should be noted that these are freely accessible data sets, they are geographically and temporally restricted; being only available for internal flights or at country level (rather than airport) travel or are usually only available at the quarterly temporal resolution (rather than monthly). Another issue is the lack of consensus regarding which data type (number of seats, passenger or flights for example) are reported by these data. This makes the potential aggregation of different data sets difficult (as they are not representing the same factors) and the comparison of models using them difficult as well. Open access data sets have the advantage of providing clear data collection methods (which is not always the case with closed source data) and can be used for validating a modelling group's data prior to its use. This additional step of validating data (as defined in **Chapter 2**) was found to be infrequently performed by researchers yet should be encouraged. Data validation provides the researcher with a clear understanding of what the data represents and identifies any errors and/or trends it may contain, as well as providing the reader with confidence that the work presented is valuable and noteworthy. Another data quality check often overlooked by researchers was the detailed reporting of the data set name, date of download, date range and any manipulation done to the data prior to use, such as grouping nearby airports under one label. As a result, a set of reporting guidelines containing all the fields required to make the use of a third-party data reproducible was generated **(Figure 2.1B)**. It is hoped that this will be used by other research groups in the future. Finally, the majority of articles selected here focused on viruses and very few on other pathogens. Although viruses like Influenza are a known threat to the global community, bacteria (including antibiotic resistant ones) are also known to be carried internationally by people, including airline passengers, yet are seldom considered for modelling.

The author conducted initial quality checks to understand what the OAG data represented in terms of passenger numbers and determine any potential trends and biases. It soon became apparent that the data showed a strong seasonality and that specific airports were shown to play specific roles. For example, ATL was predominantly used for connections whereas PEK was very few connections compared to the number of departing and arriving bookings. However, it was identified that OAG, sold as international airline data between airports (OAG, 2013) also contained railway stations as well as bus and ferry terminals among the routings provided, each with their own IATA codes. As these could not be assigned to an airport, the codes were kept in the data but their presence recorded. Recording of these stations was only noted in Bobashev G. *et al* (2008) but had not been shared with the authors in any correspondence with the company. This highlights the importance of knowing one's data as issues may not be reported but still be present when the data are bought at high cost. Additionally, by checking one's data, any collection errors lay become apparent and may be addressed appropriately if possible, as was done here.

The strong network seasonality seen in the previous chapter was linked to countries in the northern hemisphere, as this is where the majority of the global population resides. It was also noted that some countries are strongly connected to each other, such as Spain and the United Kingdom. These trends are likely to be influenced by passenger purpose of travel to these destinations, such as travelling for leisure to Spain from the UK, for example. When directly comparing open access data sets to OAG, a general overall agreement could be noted, with a large amount of noise and some discrepancies. When validating the OAG data against open access data sets to understand how many passengers were included in the OAG bookings, it was clear that in the aggregated airport level data that one booking represented roughly one passenger. However, when directly comparing airport level data, differences in how many passengers were included in each booking became apparent such that smaller airports and countries showed some discrepancies with OAG data. Because of this overall agreement, it was decided not to adjust the data and continue the analyses with OAG bookings representing one passenger.

The use of accurate airline data may provide important information regarding the international spread of human pathogens. For example, using returning airline passengers as a sentinel can help understand the changing epidemiology of disease in a given country (Fricker and Steffen, 2008; Lopez *et al.*, 2016), as well as the risk in the passenger's home country upon their return, as infected humans can transmit pathogens to local vectors (Angelini R *et al.*, 2007). When comparing returning chikungunya and dengue cases with that of returning travellers, no clear link could be identified, suggesting that the risk of infection faced by travellers was not relative to the number of passengers but rather seasonal according to their destination. Additionally, the age pattern of the returning cases provides insights into their exposure risks that should be considered by point of care clinicians. However, this airline travel data along with endemic incidence, also allowed the author to understand the risks faced by passengers compared to local populations within visited countries and also when taking duration of travel into account. When travelling to a country with a known infectious disease risk, it can be assumed that a percentage of passengers would take necessary precautions, if the risk is known prior to travel. However, sentinel surveillance also allows medical professionals to treat returning passengers quicker by knowing which pathogen to suspect first when patients return from international travel, as well as informing future travellers. The within country risk of dengue and chikungunya varied with the destination, such that UK residents travelling to the Caribbean were at reduced risk of dengue or chikungunya infection compared to the local population. On the other hand, travellers to Lower South America faced a higher risk of contracting dengue compared to the local populations. However, when including duration of travel in the model, there was an overall protective effect for travellers compared to local populations, with some variations between regions. For example, travellers to the Caribbean still faced a reduced risk for dengue but faced an increased risk of chikungunya infection in Southwest Pacific, compared to local populations. However, considering the variations in passenger number to these regions, the absolute risk showed large discrepancies with the relative risk, with North America being the safest destination according to the absolute risk whereas South East Asia was the riskiest for dengue and the Caribbean for chikungunya when using the relative risk. Passenger purpose of travel could not be included in this analysis as it was absent from the original data, however, this is very likely plays an important role in the varying levels of risks faced by travellers compared to local populations. Although reporting of dengue and especially chikungunya may not always be systematically done by countries and UK residents and doctors may not be aware of what infection they are faced with, this was the first attempt (that the author is aware of) of using sentinel data to understand within country risks faced by travellers.

Finally, global pandemic preparedness was analysed by understanding how likely an outbreak would be to spread internationally when started from a given seed country. Understanding the level of the healthcare provided in each country was attained using two indices, both combining a number of factors such as level of disease in population (GBD) or politics as well as demography and education among others (Rand Index). It was determined early that using a single factor such as measles vaccination or GDP was not representative of an HCS. Using two different indices to determine the level of healthcare provided by each country and their global connectivity, it was determined that India was the country with the potential to cause the biggest threat to global populations, followed by either Indonesia (Global Burden of Disease's Healthcare Quality Index) or Pakistan (Rand Corporation Index) depending on the index used. Additionally, the countries from which recent pandemics have spread (Brazil, China (including Hong Kong), Mexico and Saudi Arabia) showed some of the lowest risks for global pandemic spread. This is in contrast with the large amount of literature stating that countries with poor healthcare systems are more likely to see the development of an outbreak, which, with slow within-country detection has the potential to cause a pandemic. However, these groups do not also consider each country's global connectivity, except for Bogoch, I. *et al* (2018). This analysis does not mean that a vulnerable country won't be able to control an outbreak, but rather that there is a significant risk that it will generate a pandemic if an outbreak isn't controlled early (Moore *et al*. 2016). Although access to healthcare has improved globally overall, the gap between countries providing good and poor healthcare is still widening, with the majority of HCS not ready to deal with a pandemic (International Working Group on Financing Preparedness, 2017). Additionally, civil unrest, humanitarian and natural disasters have a direct and significant impact on HCS (Bonds *et al*. 2018). However, for an outbreak to develop into a pandemic the pathogen must be transmissible with relative ease and with minimal detection (for example, prior to the symptomatic phase), reach a population that can afford to fly internationally as well as coming into contact with potentially infectious people. If these events occur in a location with a vulnerable healthcare system and relatively good global connectivity, there is a high risk that an outbreak will reach other countries and potentially cause a public health event of international concern. Therefore, it is in every country's interest to enhance their pandemic preparedness, not only because it will cause fewer deaths in their country but also be financially cheaper to them and to the global community. In 2018, the WHO and World Bank co-created the Global Preparedness Monitoring Group to tackle this issue. This analysis has some limitations, including the absence of other modes of transport and the only use of historical data and the author was unaware of which index was best suited for this analysis. However, this was the first analysis of its kind at the global level.

A brief summary of the key findings described above can be found in **Box 7.1**.

Box 7.1: Overview of thesis aims and summary of key findings.

Aims of thesis:

- Understand what the airline data represents in terms of passenger movements and
- Determine whether its use was appropriate to understand the international spread of human infectious diseases.

Summary of key findings:

- **Chapter 2** showed that expensive data sets (IATA and OAG) were most often used to model the international spread of human infectious diseases and poor reporting of sources.
- Validating data (as defined in **Chapter 2**) prior to use was infrequently performed but should be encouraged.
- The author developed a reporting frame work outlining the minimum information to report when using third party data to make the work reproducible by others.
- From **Chapter 3** clear seasonal trends in travel: peaks in July-August, troughs in November, corresponding to their summer and winter months of the Northern hemisphere.
- OAG was sold as international airline data between airports, yet also contained railway, bus and ferry stations and terminals.
- The strong seasonality was driven by countries such as China and the United States. Passenger purpose of travel was also shown to play an important role in seasonality of travel.
- In **Chapter 4**, a comparison of airport-level data combined from various data sets showed that OAG bookings represent 0.91 passenger per booking. Therefore, one passenger per booking ratio was used for the rest of this thesis.
- Age and seasonality patterns can be seen in the imported cases of chikungunya and dengue from **Chapter 5**.
- The airline travel data, along with endemic prevalence and duration of travel, showed that UK passengers were at reduced risk of becoming infected with dengue when visiting the Caribbean, than the local populations. On the other hand, passengers to South Asia were at highest risk of contracting chikungunya, compared to the local population.
- Sentinel surveillance allows medical professionals to understand which potential pathogens patients returning from international travel may be infected with, as well as informing future travellers regarding the within-country risks.
- From the global pandemic preparedness analysed in **Chapter 6** it was determined that India, Pakistan and Indonesia are most likely to see the initial spread of a pandemic.
- Two indices made up of several factors were used to describe HCS but no understanding of which one is best could be determined from this analysis alone.
- There is an abundance of literature stating that poor healthcare systems lead to disease outbreaks, however the author was only aware of one group (at time of writing) who combined this information to international travel.

Limitations

The systematic literature review had the limitation of only including international travel articles, therefore, some relevant articles were knowingly excluded. As this thesis related to the international spread of infectious diseases, it was not deemed relevant to include articles relating to national travel patterns, even though these may impact the further international spread of a pathogen.

Coding and data constraints have had an impact on the level of detail that could be attained for **Chapters 5** and **6**. The risks faced by airline passengers when travelling abroad (**Chapter 5**) could only be modelled for the years 2010 to 2014 as the airline data used was restricted to those years. Therefore, it was not possible to consider the risk for Zika during the 2016 pandemic, another mosquito borne virus carried by the same vectors as the chikungunya and dengue viruses, without access to the airline travel data contemporary to the outbreak. Additionally, the transmission seasonality for chikungunya and dengue could not be taken into consideration as the endemic prevalence in each visited country was only available at the annual level. Knowing the strong role played by seasonality in the transmission of these pathogens, this is a clear limitation to the analysis that could be addressed in future work. Additionally, the absence of duration and purpose of travel from the returning passenger data, have implications on the accuracy with which the author was able to model the within-country risks. It was shown in the analysis that duration of travel impacts on the risks faced when travelling, however, this was imputed from an independent data source which may not be as accurate as if the information from the patient directly.

When considering the level of healthcare development for each country (**Chapter 6**), two indices were used, made up of several individual factors representing several aspects of the healthcare system. A clear limitation of this analysis is the absence of temporal variation for each country. In fact, between 2010 and 2015, a number of countries are likely to have seen their healthcare system quality vary as a consequence of conflict (Syria for example), important health events (Ebola in West Africa for example) or other known or unknown causes. The lack of annual values for these indices is a clear limitation as this may provide important information on the varying levels of risk countries pose to the global community according to their concurrent level of HCS development. For example, it was shown in **Chapter 6** that the connectivity of countries like Syria and West African countries significantly reduced during the conflict and the Ebola outbreak, respectively, however, according to this analysis, their healthcare system development stayed constant. This is unlikely to be true;

but, the exact extent of the variation generated by these events remains unclear. Additionally, this analysis only considers airline travel without considering land or sea travel, which is also likely to disseminate pathogens internationally. Finally, there was no breakdown of within-country regions which is problematic for countries such as India that are geographically large and may have varying baseline risks of developing an outbreak depending on the within-country region.

Implications of research

Mathematical models using airline passenger data are increasingly used to inform public health policy (Basu and Andrews, 2013). However, a number of limitations considering their use and origin need to be considered by researchers and policy makers alike. Firstly, the validity of the airline data itself. As shown in **Chapter 2**, commercial airline data, such as IATA and OAG, is most frequently used by researchers, and it can be assumed that because these data come at a very high price, they are taken as the truth. Because researchers are not second guessing the data, or at least not reporting that they are, their validity is assumed not to be questioned. However, when doing some simple analyses for this thesis, it became clear that the OAG data were not perfect, namely because they included railway stations as 'connections' in routings with recognised departing and arriving airport codes. This was not disclosed by the company and the routings could not be corrected as no airport code could not be attributed as a replacement for those railway stations. Additionally, the data were sold as airline 'Adjusted.Bookings' and 'Unadjusted.Bookings', with no indication from their documentation regarding the differences between the two, and what a booking represents in terms of passenger count. When asked in personal communications what the differences were between each type of booking and whether they could provide additional information regarding their collection methods, the company did not provide any additional details. However, they informed the author that the 'Adjusted.Bookings' were more accurate, which, when plotting against time, matched the author's *a priori* knowledge of the seasonal trends of global travel. Therefore, this thesis has provided the first in-depth description of a closed-source data set (the author is aware of), highlighting problems with the collection methods and with the data themselves.

As well as not reporting what the data represent, detailed reporting of the data set names and sources used in models is infrequently communicated by researchers. While it is increasingly requested in other fields such as biological sciences (Nature, 2014), accurate reporting of third party data, such as airline data, does not seem to be held accountable to the same reporting standards. With so many data sources available and with varying levels of quality, strongly encouraging a set of reporting standards, leading to reproducibility is in the interest of the field of research and the basis of scientific research. Additionally, the availability of a data set that the modelling community can agree upon using, providing accurate and detailed information about airline data would make the comparison of models more realistic. Until this is a reality, policy makers and journals relying or publishing these models (either to influence policy or to understand outbreak developments) should encourage researchers to report data according to the guidelines developed from this thesis and detailed in **Figure 2.1B**.

Furthermore, providing a description and performing a validation of the airline data used for these models will give readers an insight into their accuracy. As many research groups are using commercial airline data, knowing their flaws may encourage them to validate them before use. At the time of writing, the author was unaware of any such data comparisons having previously been attempted or published, and one was therefore attempted with the aim of understanding how many passengers were included in each 'Adjusted.Bookings' from OAG. Using four open access and independent data sets, it became apparent that when considering single airports and countries there were some discrepancies between the open access data sets and OAG, especially when considering small airports. Similar discrepancies were apparent when considering countries. However, when open access data sets were aggregated together to the airport or country level, the overall ratio of passenger per booking was much closer to one. Nevertheless, the OAG data severely underestimated the number of passengers departing from airport code 'USA' (Concorde airport in the United States, USDoT data) and passengers departing from Greenland (UK ONS data). **Chapter 4** aimed to provide the first validation of a commercial data set with open access data to determine the validity of the former. Additionally, it was determined that possible reasons behind these variations at the airport level of passengers per booking may also include a passenger's purpose of travel: if travelling for work, the booking is more likely to only consider one passenger, whereas if travelling for leisure or to visit friends and relatives, a booking may consider multiple passengers. These variations may also reflect the choice of departure and destination airport for passengers, such that business passengers travelling between London and Europe are more likely to use London City airport due to its proximity to the financial district (Civil Aviation Authority, 2011), for example.

To the author's knowledge, this thesis showed the first attempt at using sentinel data to understand within-country risks of chikungunya and dengue faced by UK travellers. It was determined that duration of travel within each country played an important role and overall protective effect in the level of risk encountered by the travellers. This analysis gives a first impression of the impact of duration of travel by country using aggregated data from another source collected for a different purpose (TravelPac data). Although the risk of contracting a VBD when travelling to affected regions is known to vary according to factors such as duration of travel and within country behaviour, little data was available regarding the within-country behaviours and therefore could not be included in the model. For example, passengers travelling to visit rural locations to visit relatives or for backpacking will face a different level of risk than passengers staying in air-conditioned hotel and spending their time in urban areas. However, there is little data available combining duration and purpose of travel with passenger numbers and demographics by country other than TravelPac, which is not very detailed in terms of passenger numbers as previously described. Having such detailed data freely available would be an asset to the field. In a time when the global population is increasingly well connected with airline travel, it can be hoped that the more information modellers have available to use to build their models, the more representative it will be. Indeed, this lack of information may have a detrimental effect when it comes to modelling within-country risks for varying diseases and potentially for the understanding of the spread of the next pandemic. In must also be remembered that knowledge of the risks faced by passengers according to their duration of travel and within-country behaviour will also have an impact on how travel clinics disseminate information and how clinicians treat returning passengers.

Finally, using global airline connectivity and the level of healthcare development in each country could help understand which countries may pose a higher risk in the potential initial spread of a pandemic. Such an analysis had only been previously attempted at the national level in relation to the 2017 plague outbreak in Madagascar by Bogoch *et al*. (2018), but not at the global level. An abundance of literature states that countries with poor healthcare development are more likely to see the development of an outbreak (Barber *et al*., 2017; Bonds *et al*., 2018; Elmahdawy *et al*., 2017; Moore *et al*., 2016), however, this is not often considered in combination with their international airline connectivity, which is of importance when considering the spread of pandemics. The results from this analysis showed that a different set off countries pose a higher risk to the global community (India, Pakistan and Indonesia) than those with the poorest healthcare development (Central African Republic, Somalia and South Sudan) (Moore *et al*, 2016 and International Working Group on Financial Preparedness, 2017) or those which have recently seen the development of an outbreak that developed into a pandemic (Mexico, Brazil and Saudi Arabia). Given the number of passengers who travel internationally every day, the global community should be encouraged to consider healthcare development and international travel together. This analysis has highlighted the fact that there are very few indices freely available to represent healthcare development and the factors used (for example, level of infectious diseases, education, politics…) were different in the two indices used here. As there does not seem to be a set of guidelines advising which factors may be more or less representative of healthcare development, the development of more indices using a range of factors but with comparable methods, would be helpful for the field and for future policy development at the international level. Additionally, understanding the impact of civil unrest and humanitarian crises by having a more detailed historical view at changes over time would be very useful to understand the potential future risks faced by the global community.

Understanding which countries may pose a more significant international risk provides crucial information about where to (re-)direct international aid over a longer period of time, which is more advantageous to a country than short term investment. Such aid allows the development of the country's infrastructure, thereby becoming more resilient to future disease outbreaks (Bonds *et al.*, 2018; Harvard Global Health Institute, 2018; Moore *et al.*, 2016). By assisting in the strengthening of HCS of the most vulnerable countries, the global community will become a safer place. Additionally, by providing healthcare workers in the UK information about which countries pose varying risks to travellers will help them identify potential pathogens quicker when patients present to them and disseminate the correct information from travel clinics.

Future perspectives

A number of projects could derive from this thesis, of which a few are described below. Firstly, a detailed direct comparison of the two data sources principally used by mathematical modellers, IATA and OAG, has not yet been reported, therefore it is still unclear whether one is better than the other. If this is the case, researchers and organisations relying on this data should be made aware and encouraged to use the most appropriate data set. Therefore, a continuation of this thesis would include undertaking such a comparison at the most refined level of detail possible, at least monthly airport data comparison. Although potentially very costly for the research group, it would help determine the differences between both data sets and whether one is indeed more favourable to use for infectious disease modelling than the other. This may also be of interest to the companies themselves if they wish for researchers to use their data for public health purposes.

Another important follow on would be to generate an open access data base that combines airline travel information and demographic information, such as age group, sex, size of travel party, purpose of travel and within-country activities. Such information could result from a collaboration with social scientists to gain a detailed understanding of behaviours during travel. Although TravelPac does provide some of this information, gaining more detailed data is likely to require more in-depth or different interviews than the ones being undertaken by the ONS team. This data base would need to be as accurate as possible both geographically (at airport level) and temporally (at least monthly, although the daily and/or weekly variations may also provide key insights). Given that IATA and OAG are data sources generated by and for the airline industry, generating a database specifically for epidemiological purposes, is very likely to provide key insights into the development and geographical extent of a pandemic, thus providing more accurate and more reliable information for policy makers. Such a data base wold also allow researchers to use the same information, therefore making their work more comparable between them. However, the data base would need to be updated on a regular basis to keep the information relevant to outbreak scenarios, which may require cooperation from the airline data providers themselves.

From the dengue and chikungunya analysis chapter, seasonality could not be considered due to the nature of the endemic levels of disease identified. A future piece of work could include this factor as seasonal transmission is known to be very important in the spread of these VBDs and will likely impact the relative risk for UK passengers in different regions of the world. Additionally, adding within-country activities as a model parameter could provide additional insights into which behaviour pose more or less of a risk to UK travellers and in which countries.

Understanding the impact of conflict and important events (such as natural disasters or severe outbreaks) on HCS over time would provide important additional information for international health policy makers. This would provide insights into which countries to target for long term development aid, to improve base levels of vaccination, sanitation and basic healthcare needs. Another perspective is the breakdown of the index scores by within-country regions (for the highest risk countries to start off with, such as India, Pakistan and Indonesia) and seeing the geographical variations. Accomplishing this work would require access to detailed data relating to health, and recreating the work done in the indices, which was beyond the scope of this thesis. Therefore, the creation of an open access database collating annual and sub-national data directly and indirectly relating the healthcare could provide key insights into the impact of unrest on healthcare, which country may pose varying threats and how these change over time.

Finally, media reporting of an outbreak of international concern is likely to have an impact on its containment from the speed of the international response before it escalates to a pandemic. For example, during the 2014 West African Ebola outbreak, WHO and the global community were criticised for their slow response (BBC, 2014). Indeed, international aide and press coverage were slow and minimal until cases from Western countries were identified. On the other hand, when the Zika epidemic started to affect Brazil a few months before the 2016 Rio de Janeiro Olympic Games, the response was much faster, and news coverage more important. The significant variations in reporting between these outbreaks is thought provoking (Hayden, 2016) and merits a detailed analysis regarding the factors influencing these reports and the impact media reporting may have on the development of the outbreak and pandemic.

Bibliography

2017. *GLEAMviz - The Global Epidemic and Mobility Model* [Online]. Available: http://www.gleamviz.org/ [Accessed 28th March 2017.

Abimbola, S., Topp, S. M., Palagyi, A.*, et al.* 2017. Global health security: where is the data to inform health system strengthening? *BMJ Glob Health,* 2**,** e000481.

Al-Tawfiq, J. A., Zumla, A.and Memish, Z. A. 2014. Travel implications of emerging coronaviruses: SARS and MERS-CoV. *Travel medicine and infectious disease,* 12**,** 422-428.

Al Hajjar, S.and Mcintosh, K. 2010. The first influenza pandemic of the 21st century. *Annals of Saudi Medicine,* 30**,** 1-10.

Alshammari, S. M.and Mikler, A. R. 2016. Modeling disease spread at global mass gatherings: Data requirements and challenges. 463**,** 17-26.

Amaral, L., Scala, A., Barthelemy, M.*, et al.* 2000. Classes of small-world networks. *PNAS,* 97**,** 11149-11152.

Amesh, A., Watson, M., Toner, E.*, et al.* 2018. The characteristics of pandemic pathogens. Johns Hopkins University Bloomberg School of Public Health.

Anderson, R.and May, R. 1992. *Infectious Diseases of Humans - Dynamics and Control,* Oxford, Great Britain, Oxford University Place.

Angelini R, Finarelli A, Angelini P*, et al.* 2007. An outbreak of chikungunya fever in the province of Ravenna , Italy. *Euro Surveill,* 12**,** 281.

Arcilla, M. S., Van Hattem, J. M., Haverkate, M. R.*, et al.* 2017. Import and spread of extended-spectrum β-lactamase-producing Enterobacteriaceae by international travellers (COMBAT study): a prospective, multicentre cohort study. *The Lancet Infectious Diseases,* 17**,** 78-85.

Arnal, A., Ottmann, M.and Garnier, R. 2011. Circulation des virus Influenza A: apports de la modelisation. *Virologie,* 15**,** 363-70.

Balcan, D., Colizza, V., Gonçalves, B.*, et al.* 2009. Multiscale mobility networks and the spatial spreading of infectious diseases. *Proceedings of the National Academy of Sciences of the United States of America,* 106**,** 21484-21489.

Barber, R. M., Fullman, N., Sorensen, R. J. D.*, et al.* 2017. Healthcare Access and Quality Index based on mortality from causes amenable to personal health care in 195 countries and territories, 1990–2015: a novel analysis from the Global Burden of Disease Study 2015. *The Lancet,* 390**,** 231-266.

Barrat, A., Barthelemy, M., Pastor-Satorras, R.*, et al.* 2014. The architechture of complex weighted networks. *PNAS,* 101**,** 3747–3752.

Basu, S.and Andrews, J. 2013. Complexity in mathematical models of public health policies: a guide for consumers of models. *PLoS medicine,* 10**,** e1001540.

Basundra, S., Hiremath, R. N., Khajuria, R.*, et al.* 2016. Zika virus: An emerging public health challenge. *Journal of Krishna Institute of Medical Sciences University,* 5**,** 5-12.

BBC 2014. Ebola: Kofi Annan 'bitterly disappointed' by response to Ebola.

BBC 2018. Australia-UK: first non-stop flight arrives in London from Perth.

Bernasconi, O. J., Kuenzli, E., Pires, J.*, et al.* 2016. Travelers Can Import Colistin-Resistant Enterobacteriaceae, Including Those Possessing the Plasmid-Mediated mcr-1 Gene. *Antimicrob Agents Chemother,* 60**,** 5080-4.

Bluedot.Global. *Bluedot.global* [Online]. Available: https://www.bluedot.global/about [Accessed 18th February 2018.

Bobashev, G., Morris, R. J.and Goedecke, D. M. 2008. Sampling for global epidemic models and the topology of an international airport network. *PLoS One,* 3**,** e3154.

Bogoch, I. I., Brady, O. J., Kraemer, M. U. G.*, et al.* 2016. Potential for Zika virus introduction and transmission in resource-limited countries in Africa and the Asia-Pacific region: a modelling study. *The Lancet Infectious Diseases,* 16**,** 1237-1245.

Bogoch, I. I., Creatore, M. I., Cetron, M. S., *et al.* 2015. Assessment of the potential for international dissemination of Ebola virus via commercial air travel during the 2014 west African outbreak. *Lancet,* 385, 29-35.

Bogoch, I. I., Maxim, T., Acosta, H., *et al.* 2018. Potential plague exportation from Madagascar via international air travel. *The Lancet Infectious Diseases,* 18, 247-248.

Bonds, M. H., Ouenzar, M. A., Garchitorena, A., *et al.* 2018. Madagascar can build stronger health systems to fight plague and prevent the next epidemic. *PLoS Negl Trop Dis,* 12, e0006131.

Bowen Jr, J. T.and Laroe, C. 2006. Airline networks and the international diffusion of severe acute respiratory syndrome (SARS). *Geographical Journal,* 172, 130-144.

Bozick, B. A.and Real, L. A. 2015. The Role of Human Transportation Networks in Mediating the Genetic Structure of Seasonal Influenza in the United States. *PLoS pathogens,* 11, e1004898.

Brookmeyer, R. 1991. Reconstruction and future trends of the AIDS epidemic in the United States. Science 253(5015):37-42.

Brower, V. 2003. Health is a global issue. The SARS epidemic was a wake-up call for public health authorities worldwide about the threat of emerging infectious diseases. *EMBO Reports,* 4, 649-651.

Burki, T. 2017. Plague in Madagascar. *The Lancet,* 17.

Centers for Disease Control and Prevention 2014. World Health Day 2014. *In:* CENTRE FOR DISEASE CONTROL AND PREVENTION (ed.).

Chan, E. H., Scales, D. A., Brewer, T. F., *et al.* 2013. Forecasting high-priority infectious disease surveillance regions: A socioeconomic model. *Clinical Infectious Diseases,* 56, 517-524.

Chang, C., Ortiz, K., Ansari, A., *et al.* 2016. The Zika outbreak of the 21st century. *Journal of Autoimmunity,* 68, 1-13.

Charu, V., Zeger, S., Gog, J., *et al.* 2017. Human mobility and the spatial transmission of influenza in the United States. *PLoS computational biology,* 13, e1005382.

Chen, L. H., Leder, K., Barbre, K. A., *et al.* 2018. Business travel-associated illness: a GeoSentinel analysis. *J Travel Med,* 25.

Chin, C., Sorenson, J., Harris, J., *et al.* 2011. The origin of the Haitian cholera outbreak strain. *The New England Journal of Medicine,* 364, 33-42.

Choudhury, R., Panda, S.and Singh, D. V. 2012. Emergence and dissemination of antibiotic resistance: A global problem. *Indian Journal of Medical Microbiology,* 30, 384-390.

Civil Aviation Authorities 2011. Flying on business - a study of the UK business air travel market.

Cliff, A.and Haggett, P. 2004. Time, travel and infection. *Br Med Bull,* 69, 87-99.

Clinton, C.and Sridhar, D. 2017. *Governing global health - who runs the world and why?,* New York, United States, Oxford University Press.

Cori, A., Donnelly, C. A., Dorigatti, I., *et al.* 2017. Key data for outbreak evaluation: building on the Ebola experience. *Philos Trans R Soc Lond B Biol Sci,* 372.

Cowling, B., Park, M., Fang, V., Wu, P., Leung, G. And Wu, J. 2015. Preliminary epidemiologic assessment of MERS-CoV outbreak in South Korea, May-June 2015. *Euro Surveillance,* 20, 7-13.

Dawood, F. S., Iuliano, A. D., Reed, C., *et al.* 2012. Estimated global mortality associated with the first 12 months of 2009 pandemic influenza A H1N1 virus circulation: a modelling study. *The Lancet Infectious Diseases,* 12, 687-695.

Dhama, K., Malik, Y. S., Malik, S. V. S., *et al.* 2015. Ebola from emergence to epidemic: The virus and the disease, global preparedness and perspectives. *Journal of Infection in Developing Countries,* 9, 441-455.

Doyle, M. 2014. Ebola crisis: why is the UN response taking so long? BBC International Development Correspondent, Accra.

Elmahdawy, M., Elsisi, G. H., Carapinha, J.*, et al.* 2017. Ebola Virus Epidemic in West Africa: Global Health Economic Challenges, Lessons Learned, and Policy Recommendations. *Value in Health Regional Issues,* 13**,** 67-70.
Emanuelson, U.and Egenvall, A. 2014. The data - sources and validation. *Prev Vet Med,* 113**,** 298-303.
Epstein, J. M., Goedecke, D. M., Yu, F.*, et al.* 2007. Controlling pandemic flu: The value of international air travel restrictions. *PLoS ONE,* 2.
European Centre for Disease Prevention and Control 2010. The 2009 A(H1N1) pandemic in Europe Stockholm: ECDC.
European Centre for Disease Prevention and Control 2012. Guidelines for the surveillance of invasive msoquitoes in Europe. Stockholm, Sweden.
European Centre for Disease Prevention and Control. 2013. *Epidemiological update: Outbreak of dengue in Madeira, Portugal, 13 Februrary 2013* [Online]. Available: https://ecdc.europa.eu/en/news-events/epidemiological-update-outbreak-dengue-madeira-portugal-13-february-2013 [Accessed 24 July 2017.
European Centre for Disease Prevention and Control 2014. Outbreak of Ebola virus disease in West Africa. Fourth update, 3 September 2014. *Outbreak of Ebola virus disease in West Africa.* Stockholm, Sweden: ECDC.
European Centre for Disease Prevention and Control. 2015. *Epidemiological update- Outbreak of Ebola virus disease in West Africa* [Online]. Available: http://ecdc.europa.eu/en/press/news/_layouts/forms/News_DispForm.aspx?List=8db7286c-fe2d-476c-9133-18ff4cb1b568&ID=1151 [Accessed 30th March 2017.
Ferguson, N. M., Keeling, M., Edmunds, J.*, et al.* 2003. Planning for smallpox outbreaks. *Nature,* 425**,** 681-685.
Fineberg, H. V. 2014. Pandemic preparedness and response--lessons from the H1N1 influenza of 2009. *N Engl J Med,* 370**,** 1335-42.
Finnie, T. J., South, A., Bento, A.*, et al.* 2016. EpiJSON: A unified data-format for epidemiology. *Epidemics,* 15**,** 20-6.
Fraser, C., Donnelly, C. A., Cauchemez, S.*, et al.* 2009. Pandemic potential of a strain of influenza A (H1N1): Early findings. *Science,* 324**,** 1557-1561.
Fricker, M.and Steffen, R. 2008. Travel and public health. *Journal of Infection and Public Health,* 1**,** 72-77.
Furuya-Kanamori, L., Liang, S., Milinovich, G.*, et al.* 2016. Co-distribution and co-infection of chikungunya and dengue viruses. *BMC Infect Dis,* 16**,** 84.
Gardner, L. M., Chughtai, A. A.and Macintyre, C. R. 2016. Risk of global spread of Middle East respiratory syndrome coronavirus (MERS-CoV) via the air transport network. *Journal of travel medicine,* 23.
Gautret, P., Cramer, J., Field, V.*, et al.* 2012. Infectious diseases among travellers and migrants in Europe, EuroTravelNet 2010. *Euro Surveill,* 17.
Gavi 2017. Annual Progress Report 2016. *In:* ENDEAN, C. (ed.). Geneva, Switzerland: Gavi.
Giavarina, D. 2015. Understanding Bland-Altman Analysis. *Biochemia Medica*, 25, 141-145.
Glaesser, D., Kester, J., Paulose, H.*, et al.* 2017. Global travel patterns: an overview. *J Travel Med,* 24.
Gomes, M. F., Pastore, Y. P. A., Rossi, L.*, et al.* 2014. Assessing the international spreading risk associated with the 2014 west african ebola outbreak. *PLoS Curr,* 6.
Government of India 2015. National vector borne disease control programme - Annual report 2014-2015. Ministry of health & family welfare.
Grandadam, M., Caro, V., Plumet, S.*, et al.* 2011. Chikungunya virus, southeastern France. *Emerging infectious diseases,* 17**,** 910-3.
Grassly, N. C.and Fraser, C. 2008. Mathematical models of infectious disease transmission. *Nat Rev Microbiol,* 6**,** 477-87.

Guimera, R., Mossa, S., Turtschi, A., *et al.* 2005. The worldwide air transportation network: anomalous centrality, community structure, and cities' global roles. *PNAS,* 102**,** 7794-7799.

Harvard Global Health Institute 2018. Global Monitoring of Disease Outbreak Prepardness: Preventing the Next Pandemic. Harvard University, Cambridge, Massachusetts, USA: Harvard University.

Hayden, E. 2016. Spectre of Ebola haunts Zika response. *Nature,* 531**,** 19.

Heymann, D. L., Mackenzie, J. S.and Peiris, M. 2013. SARS legacy: Outbreak reporting is expected and respected. *The Lancet,* 381**,** 779-781.

Hollingsworth, T. D., Ferguson, N. M.and Anderson, R. M. 2007. Frequent travelers and rate of spread of epidemics. *Emerging Infectious Diseases,* 13**,** 1288-1294.

Holmes, A., Rambaut, A.and Andersen, K. 2018. Pandemics: spend on surveillance, not prediction. *Nature,* 558**,** 180-182.

Holmes, A. H., Moore, L. S. P., Sundsfjord, A., *et al.* 2016. Understanding the mechanisms and drivers of antimicrobial resistance. *The Lancet,* 387**,** 176-187.

Hong Kong Airport. 2018. *Air traffic statistics* [Online]. Available: https://www.hongkongairport.com/en/the-airport/hkia-at-a-glance/fact-figures.page [Accessed 2nd May 2018.

Horton, R. 2018. Offline: UHC—one promise and two misunderstandings. *The Lancet,* 391.

Hosseini, P., Sokolow, S. H., Vandegrift, K. J., *et al.* 2010. Predictive power of air travel and socio-economic data for early pandemic spread. *PLoS ONE,* 5.

Huang, Z., Wu, X., Garcia, A. J., *et al.* 2013. An open-access modeled passenger flow matrix for the global air network in 2010. *PLoS One,* 8**,** e64317.

Huppert, A.and Katriel, G. 2013. Mathematical modelling and prediction in infectious disease epidemiology. *Clinical Microbiology and Infection,* 19**,** 999-1005.

International Working Group on Financing Preparedness 2017. From Panic and Neglect to Investing in Health Security: Financing Pandemic Preparedness at a National Level - conference edition.

Jamil, Z., Waheed, Y.and Durrani, T. Z. 2016. Zika virus, a pathway to new challenges. *Asian Pacific Journal of Tropical Medicine,* 9**,** 626-629.

Johansson, M. A. 2015. Chikungunya on the move. *Trends Parasitol,* 31**,** 43-5.

Johansson, M. A., Arana-Vizcarrondo, N., Biggerstaff, B. J., *et al.* 2011. On the treatment of Airline travelers in mathematical models. *PLoS ONE,* 6.

Jones, K. E., Patel, N. G., Levy, M. A., *et al.* 2008. Global trends in emerging infectious diseases. *Nature,* 451**,** 990-3.

Jönsson, C.and Devonish, D. 2008. Does Nationality, Gender, and Age Affect Travel Motivation? a Case of Visitors to The Caribbean Island of Barbados. *Journal of Travel & Tourism Marketing,* 25**,** 398-408.

Keeling, M.and Rohani, P. 2008. *Modeling Infectious Diseases in Humans and Animals,* Princeton, New Jersey, USA, Princeton University Press.

Keeling, M. J.and Eames, K. T. 2005. Networks and epidemic models. *J R Soc Interface,* 2**,** 295-307.

Khan, K., Bogoch, I., Brownstein, J. S., *et al.* 2014. Assessing the origin of and potential for international spread of chikungunya virus from the Caribbean. *PLoS Curr,* 6.

Ki, M. 2015. 2015 MERS outbreak in Korea: hospital-to-hospital transmission. *Epidemiol Health,* 37**,** e2015033.

Klepac, P., Kissler, S.and Gog, J. 2018. Contagion! The BBC Four Pandemic - The model behind the documentary. *Epidemics*.

Kraft, S.and Havlíková, D. 2016. Anytime? Anywhere? The seasonality of flight offers in Central Europe. *Moravian Geographical Reports,* 24.

Lawyer, G. 2016. Measuring the potential of individual airports for pandemic spread over the world airline network. *BMC Infect Dis,* 16**,** 70.

Lepelletier, D., Andremont, A., Grandbastien, B., *et al.* 2011. Risk of highly resistant bacteria importation from repatriates and travelers hospitalized in foreign countries: about the French recommendations to limit their spread. *J Travel Med,* 18**,** 344-51.
Lessler, J.and Cummings, D. A. 2016. Mechanistic Models of Infectious Disease and Their Impact on Public Health. *Am J Epidemiol,* 183**,** 415-22.
Lopez, L. F., Amaku, M., Coutinho, F. A., *et al.* 2016. Modeling Importations and Exportations of Infectious Diseases via Travelers. *Bulletin of mathematical biology,* 78**,** 185-209.
Mao, L., Wu, X., Huang, Z., *et al.* 2015. Modeling monthly flows of global air travel passengers: An open-access data resource. *Journal of Transport Geography,* 48**,** 52-60.
Mier, Y. T.-R. L., Tatem, A. J.and Johansson, M. A. 2017. Mosquitoes on a plane: Disinsection will not stop the spread of vector-borne pathogens, a simulation study. *PLoS Negl Trop Dis,* 11**,** e0005683.
Mlacker, S., Shafa, G., Aldahan, A. S., *et al.* 2016. Origin of the Zika virus revealed: a historical journey across the world. *International Journal of Dermatology,* 55**,** 1369-1372.
Mooney, H.and Newton, M. 2012. The Anatomy of a Data Citation: Discovery, Reuse, and Credit. *Journal of Librarianship and Scholarly Communication,* 1.
Moore, M., Gelfeld, B., Okunogbe, A., *et al.* 2016. Identifying future disease hot spots - Infectious disease vulnerability index. Santa Monica, California, USA: Rand Corportation.
Nature 2014. Data-access practices strengthened. *Nature,* 515**,** 312.
Newman, M. 2003. The structure and function of complex networks. *SIAM Review,* 45**,** 167-256.
Nsoesie, E. K. M., Golding N, Pigott D, Brady O, Moyes C, Johansson M, Gething P, Velayudhan R, Khan K, Hay S and Brownstein J., 2016. Global distribution and envrionmental suitability for chikunguya virus, 1952 to 2015. *Euro Surveillance,* 21.
Nuriddin, A., Jalloh, M. F., Meyer, E., *et al.* 2018. Trust, fear, stigma and disruptions: community perceptions and experiences during periods of low but ongoing transmission of Ebola virus disease in Sierra Leone, 2015. *BMJ Glob Health,* 3**,** e000410.
Nutall, I. 2014. Ebola travel: vigilance, not bans. WHO.
OAG 2013. OAG launches industry's most comprehensive air passenger data and analysis tool. Las Vegas, USA.
OAG 2015. Traffic Analyser FAQs.
OAG. 2016a. *OAG – Connecting the world of travel* [Online]. Available: http://www.oag.com/about-us [Accessed 2nd August 2016.
OAG. 2016b. *OAG history* [Online]. Available: http://www.oag.com/oag-history [Accessed 2nd August 2016.
OAG. 2016c. *Traffic Analyser* [Online]. Available: http://www.oag.com/analytics/traffic-analyser [Accessed 2nd August 2016.
OAG. 2017. *Traffic Analyser FAQs* [Online]. Available: https://www.oag.com/traffic-analyser-faq [Accessed 10th July 2017.
Office for National Statistics 2014. International Passenger Survey. Overseas Travel and Tourism.
Office for National Statistics. 2016. *Travel trends 2015* [Online]. Available: http://www.ons.gov.uk/peoplepopulationandcommunity/leisureandtourism/articles/traveltrends/2015.
Office for National Statistics. 2017. *What we do* [Online]. Available: https://www.ons.gov.uk/aboutus/whatwedo [Accessed 25th September 2017.

Okeke, I.and Edelman, R. 2001. Dissemination of antibiotic-resistant bacteria across geographic borders. *CID,* 33**,** 364-369.

Omoleke, S. A., Mohammed, I.and Saidu, Y. 2016. Ebola viral disease in West Africa: A threat to global health, economy and political stability. *Journal of Public Health in Africa,* 7**,** 27-40.

Pan American Health Organisation 2011. Cholera and post-earthquake response in Haiti. *Health Cluster Bulletin.* World Health Organization,.

Pan American Health Organization 2011. Preparedness and Response for Chikungunya Virus: Introduction in the Americas. Washington DC.

Piarroux, R., Barrais, R., Faucher, B.*, et al.* 2011. Understanding the cholera epidemic, Haiti. *Emerging infectious diseases,* 17**,** 1161-8.

Porta, M. 2008. A Dictionnary of Epidemiology. *In:* S., G. & J., L. (eds.) *A Dictionnary of Epidemiology.* New York, New York, USA: Oxford University Press, Inc,.

Public Health England. 2014. *Guidance - Chikungunya* [Online]. Available: https://www.gov.uk/guidance/chikungunya [Accessed 14th November 2017.

Public Health England 2015. Chikungunya in England, Wales and Northern Ireland: 2014. London, United Kingdom: Public Health England.

Quam, M.and Wilder-Smith, A. 2015. Importation Index of Dengue to Determine the Most Probable Origin of Importation. *J Trop Med,* 22**,** 72.

Quam, M. B., Khan, K., Sears, J.*, et al.* 2015. Estimating air travel-associated importations of dengue virus into Italy. *J Travel Med,* 22**,** 186-93.

Quam, M. B.and Wilder-Smith, A. 2016. Estimated global exportations of Zika virus infections via travellers from Brazil from 2014 to 2015. *J Travel Med,* 23.

R Core Team 2017. R: A language and environment for statistical computing. 3.4.1 ed. R Foundation for Statistical Computing, Vienna, Austria.

Read, J. M., Diggle, P. J., Chirombo, J.*, et al.* 2015. Effectiveness of screening for Ebola at airports. *The Lancet,* 385**,** 23-24.

Rezza, G., Nicoletti, L., Angelini, R.*, et al.* 2007. Infection with chikungunya virus in Italy: an outbreak in a temperate region. *The Lancet,* 370**,** 1840-1846.

Ribeiro, C. D. S., Van Roode, M. Y., Haringhuizen, G. B.*, et al.* 2018. How ownership rights over microorganisms affect infectious disease control and innovation: A root-cause analysis of barriers to data sharing as experienced by key stakeholders. *PLoS One,* 13**,** e0195885.

Rvachev, L.and Longini Jr, I. M. 1985. A mathematical model for the global spread of influenza. *Mathematical Biosciences,* 75**,** 3-22.

Saker, L., Lee, K., Cannito, B.*, et al.* 2004. Globalization and infectious diseases: a review of the linkages. *Special Topics in Social, Economic and Behaviour (SEB) Research.* World Health Organisation - Geneva: TDR Steering Committee for Social, Economic and Behavioural Research.

San Martin, J. L., Brathwaite, O., Zambrano, B.*, et al.* 2010. The epidemiology of dengue in the americas over the last three decades: a worrisome reality. *Am J Trop Med Hyg,* 82**,** 128-35.

Schlangenhauf, P., Poumerol, G. And Santos-O'connor F. 2011. Microbes on the move: prevention, curtailment, oubtreak. *In:* PETERSEN, E., CHEN, L. AND SCHLAGENHAUF. (ed.) *Infectious Diseases: a geographic guide.* Wiley-Blackwell.

Semenza, J. C., Lindgren, E., Balkanyi, L.*, et al.* 2016a. Determinants and drivers of infectious disease threat events in europe. *Emerging Infectious Diseases,* 22**,** 581-589.

Semenza, J. C., Rocklov, J., Penttinen, P.*, et al.* 2016b. Observed and projected drivers of emerging infectious diseases in Europe. *Ann N Y Acad Sci,* 1382**,** 73-83.

Semenza, J. C., Sudre, B., Miniota, J.*, et al.* 2014. International Dispersal of Dengue through Air Travel: Importation Risk for Europe. *PLoS Neglected Tropical Diseases,* 8.

Sinha, M. S.and Parmet, W. E. 2016. The Perils of Panic: Ebola, HIV, and the Intersection of Global Health and Law. *Am J Law Med,* 42**,** 223-255.

Stanaway, J., Shepard,D., Undurraga, E., Halasa, Y., Coffeng, L., Brady, O., Hay, S., Bedi, N., Bensenor, I., Castaneda-Orjuela, C., Chuang, T-W., Gibney, K., Memish, Z., Rafay, A., Ukwaja, K., Yonemoto, N. And Murray, C. 2016. The global burden of dengue: an analysis from the Global Burden of Disease Study 2013. *The Lancet,* 16**,** 712-723.
Stenberg, K., Hanssen, O., Edejer, T. T.-T.*, et al.* 2017. Financing transformative health systems towards achievement of the health Sustainable Development Goals: a model for projected resource needs in 67 low-income and middle-income countries. *The Lancet Global Health,* 5**,** e875-e887.
Su, S., Wong, G., Liu, Y.*, et al.* 2015. MERS in South Korea and China: A potential outbreak threat? *The Lancet,* 385**,** 2349-2350.
Tatem, A. J.and Hay, S. I. 2007. Climatic similarity and biological exchange in the worldwide airline transportation network. *Proceedings of the Royal Society B: Biological Sciences,* 274**,** 1489-1496.
Tatem, A. J., Hay, S. I.and Rogers, D. J. 2006b. Global traffic and disease vector dispersal. *Proc Natl Acad Sci U S A,* 103**,** 6242-7.
The World Bank. 2017. *Air transport, passengers carried* [Online]. Available: http://data.worldbank.org/indicator/IS.AIR.PSGR [Accessed 7th July 2017.
Tian, H., Sun, Z., Faria, N. R.*, et al.* 2017. Increasing airline travel may facilitate co-circulation of multiple dengue virus serotypes in Asia. *PLoS Negl Trop Dis,* 11**,** e0005694.
Tourism Intelligence International 2010. How The Brits Will Travel 2010.
Tran, Q. H. A.and Namatame, A. 2016. Worldwide aviation network vulnerability analysis: a complex network approach. *Evolutionary and Institutional Economics Review,* 12**,** 349-373.
United Kingdom Government. 2017. *Guidance - Pandemic flu* [Online]. [Accessed 27th March 2018.
United Nations 2017. The Sustainable Development Goals Report 2017. New York, United States: United Nations.
United Nations Development Programme 2017. A socio-economic impact assessment of the Zika virus in Latin America and the Caribbean: with a focus on Brazil, Colombia and Suriname. the United States of America: United Nations Development Programme.
United Nations News Centre 2017. UN 'will walk with Haiti' on path towards sustainable development, senior official pledges.
US Department of Transport - Bureau of Transportation Statistics 2015. Summary 2014 U.S.-Based Airline Traffic Data.
Venkatramanan, S., Lewis, B., Chen, J.*, et al.* 2018. Using data-driven agent-based models for forecasting emerging infectious diseases. *Epidemics,* 22**,** 43-49.
Vilkman, K., Pakkanen, S. H., Laaveri, T.*, et al.* 2016. Travelers' health problems and behavior: prospective study with post-travel follow-up. *BMC Infect Dis,* 16**,** 328.
Walton, D.and Ivers, L. 2011. Responding to Cholera in Post-Earthquake Haiti. *The New England Journal of Medicine,* 364.
Wandelt, S.and Sun, X. 2015. Evolution of the international air transportation country network from 2002 to 2013. *Transportation Research Part E: Logistics and Transportation Review,* 82**,** 55-78.
Watts, D.and Strogatz, S. 1998. Collective dynamics of 'small-world' networks. *Nature,* 393**,** 440-442.
Wilder-Smith, A. 2006. The severe acute respiratory syndrome: Impact on travel and tourism. *Travel medicine and infectious disease,* 4**,** 53-60.
Wilson, M. 1995. Travel and the Emergence of Infectious Diseases. *Emerging infectious diseases,* 1**,** 39-46.
Woolley-Meza, O., Thiemann, C., Grady, D.*, et al.* 2011. Complexity in human transportation networks: a comparative analysis of worldwide air transportation and global cargo-ship movements. *The European Physical Journal B,* 84**,** 589-600.

World Bank 2017a. Drug-resistant infections: a threat to our economic future. Washington DC: World Bank.
World Bank 2017b. Population ages 0-14 (% of total), India. *In:* PROSPECTS, U. N. P. D. W. P. (ed.).
World Health Organization 2001. Weekly epidemiological record. World Health Organisation,.
World Health Organization 2003. Consensus document on the epidemiology of severe acute respiratory syndrome (SARS). Geneva, Switzerland: World Health Organization,.
World Health Organization. 2010. *Pandemic (H1N1) 2009 - update 112* [Online]. Available: http://www.who.int/csr/don/2010_08_06/en/ [Accessed 9th April 2018.
World Health Organization 2012. *International travel and health*, World Health Organisation,.
World Health Organization 2014. A global brief on vector-borne diseases. Geneva, Switzerland.
World Health Organization 2015a. 2015 Global Reference List of 100 Core Health Indicators. World Health Organization.
World Health Organization 2015b. World Health Statistics 2015 - Indicator compendium. World Health Organization,.
World Health Organization. 2016a. *Dengue and severe dengue - Fact sheet* [Online]. Available: http://www.who.int/mediacentre/factsheets/fs117/en/# [Accessed 22nd August 2016.
World Health Organization 2016b. Ebola situation report. *Ebola situation report.*
World Health Organization. 2016a. *Zika virus - Fact sheet* [Online]. Available: http://www.who.int/mediacentre/factsheets/zika/en/ [Accessed 3rd November 2016.
World Health Organization 2017a. Chikungunya - Fact sheet.
World Health Organization. 2017b. *Global EWARS* [Online]. Available: http://ewars-project.org/ [Accessed 2nd May 2018.
World Health Organization 2017c. Measles outbreaks across Europe threaten progress towards elimination. Copenhagen.
World Health Organization. 2017d. *Middle East respiratory syndrome coronavirus (MERS-CoV) - Factsheet* [Online]. Available: http://www.who.int/mediacentre/factsheets/mers-cov/en/ [Accessed 4th January 2018.
World Health Organization. 2017e. *Plague* [Online]. Available: http://www.who.int/mediacentre/factsheets/fs267/en/ [Accessed 4th January 2018.
World Health Organization. 2017f. *Plague - Madagascar* [Online]. Available: http://www.who.int/csr/don/27-november-2017-plague-madagascar/en/ [Accessed 4th January 2018.
World Health Organization. 2017g. *Vector-borne diseases* [Online]. Available: http://www.who.int/mediacentre/factsheets/fs387/en/ [Accessed 14th November 2017.
World Health Organization 2018a. Ebola virus disease - Democratic Republic of the Congo. World Health Organization.
World Health Organization 2018b. Managing epidemics: key facts about major deadly diseases. Geneva, Switzerland: World Health Organization.
World Health Organization. 2018c. *Middle East respiratory syndrome coronavirus (MERS-CoV) - Oman* [Online]. Available: http://www.who.int/csr/don/15-march-2018-mers-oman/en/ [Accessed 10th April 2018.
World Health Organization 2018d. Nipah virus - India. *Outbreak Disease News.*

World Health Organization 2018e. WHO and World Bank Group join forces to strengthen global health security. Geneva, Switzerland.

World Health Organization 2018f. WHO supports Ebola vaccination of high risk populations in the Democratic Republic of the Congo. Geneva, Switzerland.

World Health Organization 2018g. World health statistics 2018: monitoring heatlh for the SDGs, sustainable development goals. Geneva, Switzerland: World Health Organization.

World Health Organization Africa 2017. Plague outbreak Madagascar. *External Situation report.* World Health Organization Office for Africa.

World Tourism Organisation 2016. Exports from international tourism rise 4% in 2015. *UNWTO World Tourism Barometer.*

Yale Law School Roundtable on Data and Code Sharing 2010. Reproducible research - Addressing the need for data and code sharing in computational science. *Computing in Science and Engineering*, 8-12.

Appendices

Table 1: List of countries names (according to OAG) included in each OAG region and sub-region.

Region name	Sub-region name	Countries included
Africa	AF1	Algeria, Egypt, Libya, Morocco, Sudan, Tunisia
	AF2	Angola, Botswana, Lesotho, Malawi, Mozambique, Namibia, South Africa, Swaziland, Zambia, Zimbabwe
	AF3	Benin, Burkina Faso, Cameroon, Cape Verde, Central African Republic, Chad, Congo, Congo Democratic Republic of, Cote D'Ivoire, Equatorial Guinea, Gabon, Gambia, Ghana, Guinea, Guinea-Bissau, Liberia, Mali, Mauritania, Mayotte, Niger, Nigeria, Saint Helena, Sao Tome and Principe, Senegal, Sierra Leone, Togo, Zaire
	AF4	Burundi, Comoros, Djibouti, Eritrea, Ethiopia, Kenya, Madagascar, Mauritius, Reunion, Rwanda, Seychelles, Somalia, South Sudan, Tanzania United Republic of, Uganda
Asia	AS1	Afghanistan, Bangladesh, India, Maldives, Nepal, Pakistan, Sri Lanka
	AS2	Bhutan, Kazakhstan, Kazakhstan, Kyrgyzstan, Tajikistan, Turkmenistan, Uzbekistan
	AS3	Brunei Darussalam, Cambodia, Cocos (keeling) Islands, East Timor, East Timor, Indonesia, Lao People's Democratic Republic, Malaysia, Myanmar, Philippines, Singapore, Thailand, Timor-Leste, Viet Nam
	AS4!China	Chinese Taipei, Hong Kong, Hong Kong (sar) China, Japan, Korea Democratic People's Republic of, Korea Republic of, Macao (sar) China, Macau, Mongolia, Taiwan Province of China
	China	China
Europe	EU2	Albania, Armenia, Azerbaijan, Belarus, Bosnia and Herzegovina, Bulgaria, Croatia, Czech Republic, Estonia, Georgia, Hungary, Latvia, Lithuania, Macedonia Former Yugoslav Republic of, Moldova Republic of, Montenegro, Poland, Romania, Russian Federation, Serbia, Serbia and Montenegro, Slovakia, Slovenia, Ukraine and Yugoslavia
	W_EU*	Faroe Islands, France, Great Britain, Iceland, Ireland, Monaco
	South_EU*	Gibraltar, Italy, Malta, Portugal, Spain
	Scandinavia*	Denmark, Finland, Netherlands, Norway, Sweden
	Central_EU*	Austria, Belgium, Germany, Luxembourg, Switzerland
	SE_EU*	Cyprus, Greece, Turkey
Latin America	LA1	Anguilla-Leeward Islands, Antigua and Barbuda-Leeward Islands, Aruba, Bahamas, Barbados, Bermuda, Bonaire, Saint Eustatius and Saba, Cayman Islands, Cuba, Curacao, Dominica, Dominican Republic, Grenada, Windward Islands, Guadeloupe, Haiti, Jamaica, Martinique, Montserrat-Leeward Islands, Netherlands Antilles, Puerto Rico, Saint Barthelmy, Saint Kitts and Nevis-Leeward Islands, Saint Lucia, Saint Martin, St Maarten (Dutch Part), St Vincent and the Grenadines, Trinidad and Tobago, Turks and Caicos Islands, Virgin Islands-British, Virgin Islands-US
	LA2	Belize, Costa Rica, El Salvador, Guatemala, Honduras, Mexico, Nicaragua, Panama
	LA3	Bolivia, Colombia, Ecuador, French Guiana, Guyana, Peru, Suriname, Venezuela
	LA4	Argentina, Brazil, Chile, Falkland Islands, Paraguay, Uruguay
Middle East	ME1	Bahrain, Iran Islamic Republic of, Iraq, Israel, Jordan, Kuwait, Lebanon, Oman, Palestine, Qatar, Saudi Arabia, Syrian Arab Republic, United Arab Emirates, Yemen

(Table 1 continues on next page)

(Table 1 continued)

North America	Central_USA*	Arkansas, Colorado, Kansas, Louisiana, Nebraska, New Mexico, North Dakota, South Dakota, Oklahoma and Texas
	NA1!USA	Canada, Greenland, Saint Pierre and Miquelon, United States Minor Outlying Islands, USA
	NE_USA*	Connecticut, Delaware, Maine, Massachusetts, Maryland, New Hampshire, New Jersey, New York, Pennsylvania, Rhode Island, Vermont, Virginia, West Virginia
	North_USA*	Illinois, Indiana, Iowa, Kentucky, Michigan, Minnesota, Missouri, Ohio, Wisconsin
	NW_USA*	Alaska, Idaho, Montana, Oregon, Washington, Wyoming
	SE_USA*	Alabama, Florida, Georgia, Mississippi, North Carolina, South Carolina, Tennessee
	SW_USA*	Arizona, California, Hawaii, Nevada, Utah
Southwest Pacific	SW1	American Samoa, Australia, Christmas Island-Indian Ocean, Cook Islands, Fiji, French Polynesia, Guam, Kiribati, Marshall Islands, Micronesia Federated States of, Nauru, New Caledonia, New Zealand, Niue, Norfolk Island, Northern Mariana Islands (except Guam), Palau, Papua New Guinea, Samoa, Solomon Islands, Tonga, Tuvalu, Vanuatu and Wallis, Futuna Islands

* User defined sub-region grouping.

Table 4.3: Random slope coefficient values for each airport present in the open access and OAG data, with their respective number of passengers and bookings and the ratios calculated as passengers per bookings. Airport codes ending in "_dom" and "_int" represent airports in the PANYNJ data, as per **Table 4.1**.
Note: the quartile groups 1 to 4 are those shown in **Figure 4.7**.

Airport code	Airport name	Country	Random slope	Quantile group	Bookings	Passengers	Ratio
MKK	Hoolehua	USA	0.07	1	292,023	18,190	0.06
LNY	Lanai City	USA	0.11	1	259,647	27,540	0.11
ENA	Kenai	USA	0.15	1	431,033	72,680	0.17
GST	Gustavus Airport	USA	0.17	1	52,656	10,720	0.20
JHM	Kapalua	USA	0.19	1	125,854	20,530	0.16
PHO	Point Hope	USA	0.22	1	499	110	0.22
OOK	Toksook Bay	USA	0.22	1	142	20	0.14
CYF	Chefornak	USA	0.22	1	1,785	170	0.10
SPN	Saipan	Northern Mariana Islands (except Guam)	0.26	1	195,297	44,330	0.23
KKI	Akiachak	USA	0.27	1	236	60	0.25
HOM	Homer	USA	0.31	1	129,216	35,240	0.27
ROP	Rota	Northern Mariana Islands (except Guam)	0.33	1	11,991	1,420	0.12
YAK	Yakutat	USA	0.36	1	95,159	38,440	0.40
VQS	Vieques	Puerto Rico	0.37	1	107,743	32,510	0.30
VDZ	Valdez	USA	0.38	1	56,993	17,570	0.31
AIN	Wainwright	USA	0.39	1	380	150	0.39
KSM	St Mary's	USA	0.40	1	3,136	270	0.09
NUI	Nuiqsut	USA	0.40	1	115	40	0.35
MWH	Moses Lake Grant County Apt	USA	0.43	1	1,063	420	0.40
PIP	Pilot Point Airport	USA	0.43	1	55	20	0.36
VAK	Chevak	USA	0.44	1	176	80	0.45
AKN	King Salmon	USA	0.44	1	110,806	53,950	0.49
SGY	Skagway	USA	0.46	1	211	90	0.43
SNP	St Paul Island	USA	0.48	1	6,116	2,620	0.43
DLG	Dillingham	USA	0.48	1	85,417	39,010	0.46

(Table 4.3 continues on next page)

(Table 4.3 continued)

Airport code	Airport name	Country	Random slope	Quantile group	Bookings	Passengers	Ratio
PGA	Page	USA	0.50	1	24,341	12,660	0.52
ADK	Adak Island	USA	0.51	1	18,895	10,490	0.56
WRG	Wrangell	USA	0.52	1	89,691	51,620	0.58
CDV	Cordova Merle K (Mudhole) Smith Apt	USA	0.52	1	109,814	64,550	0.59
TNK	Tununak	USA	0.55	1	40	20	0.50
DUT	Dutch Harbor	USA	0.55	1	152,373	87,920	0.58
CDB	Cold Bay	USA	0.56	1	9,272	4,990	0.54
OXF	Oxford	United Kingdom	0.56	1	4,514	2,509	0.56
GUM	Guam Antonio B Won Pat International	Guam	0.57	1	451,330	282,620	0.63
GAM	Gambell	USA	0.58	1	37	40	1.08
MBL	Manistee	USA	0.58	1	10,425	4,500	0.43
CBG	Cambridge	United Kingdom	0.58	1	43,347	26,641	0.61
DDC	Dodge City	USA	0.59	1	26,607	17,120	0.64
DEC	Decatur	USA	0.60	1	5,415	1,680	0.31
MCG	Mcgrath	USA	0.60	1	2,222	780	0.35
MVY	Martha's Vineyard	USA	0.61	1	176,289	127,580	0.72
EAR	Kearney	USA	0.61	1	58,573	40,100	0.68
ACK	Nantucket	USA	0.61	1	343,417	242,180	0.71
HYS	Hays	USA	0.62	1	49,215	32,630	0.66
CEZ	Cortez	USA	0.62	1	31,102	21,890	0.70
HNS	Haines	USA	0.62	1	192	140	0.73
FMN	Farmington	USA	0.63	1	58,877	40,160	0.68
UNK	Unalakleet	USA	0.63	1	4,013	670	0.17
ADQ	Kodiak Apt	USA	0.63	1	318,708	225,830	0.71
DIK	Dickinson	USA	0.64	1	175,821	140,380	0.80

(Table 4.3 continues on next page)

(Table 4.3 continues)

Airport code	Airport name	Country	Random slope	Quantile group	Bookings	Passengers	Ratio
SHR	Sheridan	USA	0.67	1	66,989	48,280	0.72
ISN	Williston	USA	0.67	1	355,764	300,470	0.84
ART	Watertown (US) NY	USA	0.68	1	99,896	77,710	0.78
SCC	Prudhoe Bay/Deadhorse	USA	0.68	1	209,904	155,820	0.74
SDP	Sand Point	USA	0.68	1	14,748	6,540	0.44
PSG	Petersburg	USA	0.69	1	104,751	80,540	0.77
WYS	West Yellowstone	USA	0.69	1	38,916	27,660	0.71
ECP	Panama City Nw Florida Beaches Intl	USA	0.69	1	2,523,009	1,918,840	0.76
LBL	Liberal	USA	0.69	1	24,481	18,700	0.76
RIW	Riverton	USA	0.71	1	67,035	50,640	0.76
BKG	Branson	USA	0.72	1	571,570	466,180	0.82
SAF	Santa Fe (US)	USA	0.72	1	345,955	277,890	0.80
FOE	Topeka Forbes AFB	USA	0.72	1	13,790	10,790	0.78
LEB	Lebanon	USA	0.73	1	28,529	22,000	0.77
LWB	Lewisburg	USA	0.73	1	59,629	51,390	0.86
HHH	Hilton Head Island	USA	0.73	1	392,905	319,370	0.81
TUP	Tupelo	USA	0.74	1	43,497	37,750	0.87
OGD	Ogden Hinckley Apt	USA	0.75	1	34,364	30,690	0.89
ROA	Roanoke	USA	0.75	1	1,591,399	1,335,810	0.84
PIR	Pierre	USA	0.76	1	63,817	54,500	0.85
PVU	Provo	USA	0.76	1	161,366	142,260	0.88
KTN	Ketchikan International Apt	USA	0.77	1	540,702	468,950	0.87
PVC	Provincetown	USA	0.77	1	35,372	15,960	0.45
LAW	Lawton/Fort Sill	USA	0.77	1	364,298	305,140	0.84
GRK	Killeen/Fort Hood Regional/R. Gray AAF	USA	0.77	1	1,010,434	869,990	0.86

(Table 4.3 continues on next page)

(Table 4.3 continued)

Airport code	Airport name	Country	Random slope	Quantile group	Bookings	Passengers	Ratio
CRW	Charleston (US) WV	USA	0.77	1	1,366,256	1,184,210	0.87
ERI	Erie	USA	0.77	1	573,071	494,260	0.86
GYY	Chicago Gary International Apt	USA	0.77	1	17,430	14,330	0.82
VEL	Vernal	USA	0.78	1	31,748	26,030	0.82
RKD	Rockland	USA	0.78	1	24,725	15,330	0.62
CGI	Cape Girardeau	USA	0.78	1	23,202	17,390	0.75
MHK	Manhattan	USA	0.78	1	328,192	286,020	0.87
ITH	Ithaca	USA	0.79	1	579,734	508,010	0.88
GRI	Grand Island	USA	0.79	1	272,838	241,730	0.89
ATW	Appleton	USA	0.79	1	1,229,054	1,081,500	0.88
BHB	Bar Harbor	USA	0.79	1	49,287	41,850	0.85
HDN	Hayden	USA	0.79	1	540,373	514,120	0.95
AMA	Amarillo Rick Husband Intl Apt	USA	0.79	1	2,132,255	1,887,150	0.89
JLN	Joplin	USA	0.79	1	138,150	118,910	0.86
BFF	Scottsbluff	USA	0.79	1	39,046	34,140	0.87
CAE	Columbia Metropolitan Apt	USA	0.80	1	2,478,105	2,213,110	0.89
PHF	Newport News	USA	0.80	1	1,906,988	1,741,270	0.91
DAL	Dallas/Fort Worth Dallas Love Field	USA	0.80	1	16,980,240	15,251,670	0.90
FAI	Fairbanks International Apt	USA	0.80	1	1,861,681	1,657,740	0.89
GCK	Garden City	USA	0.80	1	98,169	94,290	0.96
CHO	Charlottesville	USA	0.81	1	1,145,651	1,032,640	0.90
MDT	Harrisburg International Apt	USA	0.81	1	3,221,666	2,911,570	0.90
BKW	Beckley	USA	0.81	1	17,266	13,620	0.79
GRB	Green Bay	USA	0.81	1	1,498,035	1,358,430	0.91
ELP	El Paso International Apt	USA	0.81	1	7,096,254	6,448,020	0.91

(Table 4.3 continues on next page)

(Table 4.3 continued)

Airport code	Airport name	Country	Random slope	Quantile group	Bookings	Passengers	Ratio
XNA	Fayetteville/Springdale NW Arkansas Reg	USA	0.81	1	2,784,420	2,533,420	0.91
LCK	Columbus Rickenbacker Apt	USA	0.81	1	76,512	74,880	0.98
HTS	Huntington	USA	0.82	1	571,447	521,530	0.91
TYS	Knoxville	USA	0.82	1	4,040,561	3,684,450	0.91
HSV	Huntsville International Airport	USA	0.82	1	2,847,483	2,594,780	0.91
FNL	Fort Collins/Loveland Municipal Apt	USA	0.82	1	97,819	90,240	0.92
MDW	Chicago Midway Apt	USA	0.82	1	32,227,040	29,515,740	0.92
STL	St Louis Lambert Intl Apt	USA	0.82	1	25,494,269	23,345,920	0.92
LBF	North Platte	USA	0.82	1	34,640	30,880	0.89
JNU	Juneau	USA	0.82	1	1,158,134	1,053,580	0.91
GSO	Greensboro/High Point	USA	0.82	1	4,173,342	3,830,440	0.92
PKB	Parkersburg/Marietta	USA	0.82	1	37,733	34,170	0.91
SFB	Orlando Sanford International Airport	USA	0.82	1	3,216,638	2,996,840	0.93
TOL	Toledo Express Apt	USA	0.83	1	388,257	360,180	0.93
YNG	Youngstown	USA	0.83	1	199,120	188,160	0.94
RAP	Rapid City Regional Apt	USA	0.83	1	1,290,009	1,193,910	0.93
LEX	Lexington Blue Grass Apt	USA	0.83	1	2,602,527	2,412,130	0.93
SHV	Shreveport Regional Apt	USA	0.83	1	1,399,691	1,297,870	0.93
ANC	Anchorage Ted Stevens Intl Apt	USA	0.83	1	8,047,246	7,442,000	0.92
SBN	South Bend	USA	0.83	1	1,511,994	1,401,680	0.93
MSN	Madison (US) WI	USA	0.83	1	3,779,528	3,508,810	0.93
ABQ	Albuquerque	USA	0.83	1	12,141,184	11,275,290	0.93
OTZ	Kotzebue	USA	0.83	1	150,703	139,600	0.93
HOU	Houston William P. Hobby Apt	USA	0.83	1	20,038,823	18,611,620	0.93
OGG	Kahului	USA	0.83	1	11,429,179	10,578,010	0.93

(Table 4.3 continues on next page)

(Table 4.3 continued)

Airport code	Airport name	Country	Random slope	Quantile group	Bookings	Passengers	Ratio
MLI	Moline	USA	0.83	1	1,965,234	1,827,110	0.93
RFD	Chicago Rockford Airport	USA	0.83	1	409,365	379,610	0.93
TRI	Tri-Cities Regional	USA	0.83	1	1,071,900	997,760	0.93
IAG	Niagara Falls	USA	0.83	1	328,475	315,820	0.96
SIT	Sitka	USA	0.83	1	317,134	295,910	0.93
HVN	New Haven	USA	0.83	1	210,032	193,820	0.92
AZA	Phoenix Mesa Gateway Airport	USA	0.83	1	2,632,689	2,421,940	0.92
DAY	Dayton Intl Apt	USA	0.83	1	5,975,615	5,574,910	0.93
OWB	Owensboro	USA	0.84	1	104,083	96,780	0.93
ORH	Worcester	USA	0.84	1	85,869	80,120	0.93
LIT	Little Rock	USA	0.84	2	5,326,653	4,982,990	0.94
CVG	Cincinnati Northern Kentucky Intl Apt	USA	0.84	2	9,955,490	9,320,650	0.94
AUG	Augusta	USA	0.84	2	27,906	21,920	0.79
AZO	Kalamazoo	USA	0.84	2	616,101	577,280	0.94
CID	Cedar Rapids	USA	0.84	2	2,327,840	2,187,460	0.94
ROC	Rochester (US) NY	USA	0.84	2	5,731,577	5,383,080	0.94
ICT	Wichita Dwight D. Eisenhower Apt	USA	0.84	2	3,505,782	3,304,410	0.94
PIE	Tampa St Petersbrg-Clearwater Intl Apt	USA	0.84	2	2,209,433	2,108,260	0.95
MEM	Memphis International Apt	USA	0.84	2	8,137,382	7,664,410	0.94
MOB	Mobile Municipal Apt	USA	0.84	2	1,391,357	1,310,600	0.94
TLH	Tallahassee	USA	0.84	2	1,521,613	1,433,540	0.94
FLO	Florence	USA	0.84	2	392,768	370,220	0.94
BGM	Binghamton	USA	0.84	2	530,688	501,130	0.94
SYR	Syracuse	USA	0.84	2	4,717,221	4,458,060	0.95
PGD	Punta Gorda (US)	USA	0.85	2	732,713	720,650	0.98

(Table 4.3 continues on next page)

(Table 4.3 continued)

Airport code	Airport name	Country	Random slope	Quantile group	Bookings	Passengers	Ratio
FAY	Fayetteville Municipal Apt	USA	0.85	2	1,287,845	1,219,830	0.95
MGM	Montgomery Dannelly Field	USA	0.85	2	910,502	861,630	0.95
CYS	Cheyenne Regional Apt	USA	0.85	2	76,733	71,390	0.93
JFK	New York J F Kennedy International Apt	USA	0.85	2	46,340,395	43,874,250	0.95
EVV	Evansville	USA	0.85	2	810,471	768,040	0.95
SGF	Springfield (US) MO	USA	0.85	2	1,767,571	1,678,980	0.95
LSE	La Crosse	USA	0.85	2	480,924	453,560	0.94
CHA	Chattanooga Lovell Field Apt	USA	0.85	2	1,613,807	1,535,220	0.95
STX	St Croix Henry E. Rohlsen Apt	Virgin Islands, US	0.85	2	683,620	648,360	0.95
HPN	Westchester County	USA	0.85	2	4,473,395	4,263,400	0.95
PIA	Peoria	USA	0.85	2	1,337,672	1,277,180	0.95
SCE	State College	USA	0.85	2	681,573	650,410	0.95
BWI	Baltimore Washington International Apt	USA	0.85	2	40,326,358	38,556,220	0.96
RHI	Rhinelander	USA	0.85	2	96,608	97,310	1.01
AEX	Alexandria International Apt	USA	0.85	2	681,005	650,350	0.95
RIC	Richmond (US)	USA	0.86	2	7,627,200	7,306,470	0.96
JAN	Jackson-evers International Airport	USA	0.86	2	2,918,738	2,790,820	0.96
SDF	Louisville International	USA	0.86	2	7,678,670	7,355,170	0.96
BHM	Birmingham	USA	0.86	2	6,493,202	6,220,420	0.96
SCK	Sacramento Stockton Metropolitan	USA	0.86	2	318,802	308,560	0.97
BTR	Baton Rouge	USA	0.86	2	1,922,006	1,843,760	0.96
MWA	Marion	USA	0.86	2	32,596	25,090	0.77
ORF	Norfolk International Apt	USA	0.86	2	7,452,003	7,186,950	0.96
SBY	Salisbury-Ocean City	USA	0.86	2	350,056	334,550	0.96
MLU	Monroe	USA	0.86	2	550,400	529,370	0.96

(Table 4.3 continues on next page)

(Table 4.3 continued)

Airport code	Airport name	Country	Random slope	Quantile group	Bookings	Passengers	Ratio
BUF	Buffalo	USA	0.86	2	12,170,421	11,740,790	0.96
ABE	Allentown/Bethlehem/Easton	USA	0.86	2	1,809,320	1,751,470	0.97
ELM	Elmira/Corning	USA	0.86	2	721,053	694,360	0.96
EGE	Vail/Eagle County Rgnl Apt	USA	0.86	2	821,452	832,050	1.01
OAJ	Jacksonville (US) NC	USA	0.86	2	898,836	868,580	0.97
FWA	Fort Wayne Baer Field	USA	0.86	2	1,409,350	1,359,780	0.96
DSM	Des Moines	USA	0.86	2	4,687,385	4,532,040	0.97
FSM	Fort Smith (US)	USA	0.86	2	443,129	426,090	0.96
PBG	Plattsburgh	USA	0.86	2	553,714	538,470	0.97
LRD	Laredo	USA	0.86	2	507,586	489,800	0.96
OKC	Oklahoma City Will Rogers Apt	USA	0.86	2	8,332,767	8,052,740	0.97
BMI	Bloomington-Normal	USA	0.86	2	1,183,809	1,144,590	0.97
OXR	Oxnard/Ventura	USA	0.86	2	1,982	1,860	0.94
TUL	Tulsa International Apt	USA	0.86	2	6,364,106	6,158,410	0.97
LAN	Lansing	USA	0.87	2	861,658	836,560	0.97
GRR	Grand Rapids	USA	0.87	2	5,052,674	4,885,460	0.97
CSG	Columbus Metropolitan Apt	USA	0.87	2	296,519	287,260	0.97
MCI	Kansas City International Apt	USA	0.87	2	22,036,969	21,342,820	0.97
MSP	Minneapolis/St Paul International Apt	USA	0.87	2	39,320,247	38,034,060	0.97
KOA	Kona	USA	0.87	2	5,678,751	5,490,290	0.97
PIT	Pittsburgh International Apt	USA	0.87	2	17,428,010	16,952,620	0.97
LGA	New York La Guardia Apt	USA	0.87	2	55,430,852	53,971,160	0.97
LFT	Lafayette Regional Apt	USA	0.87	2	1,080,167	1,050,940	0.97
CRP	Corpus Christi International Apt	USA	0.87	2	1,646,106	1,602,250	0.97
COS	Colorado Springs Municipal	USA	0.87	2	3,579,762	3,487,910	0.97

(Table 4.3 continued on next page)

(Table 4.3 continues)

Airport code	Airport name	Country	Random slope	Quantile group	Bookings	Passengers	Ratio
BNA	Nashville	USA	0.87	2	19,981,783	19,488,500	0.98
RST	Rochester (US) MN	USA	0.87	2	547,421	533,060	0.97
GPT	Gulfport/Biloxi	USA	0.87	2	1,466,847	1,431,990	0.98
IAD	Washington Dulles International Apt	USA	0.87	2	18,464,556	17,880,610	0.97
IND	Indianapolis	USA	0.87	2	16,419,175	16,022,060	0.98
SPS	Wichita Falls Municipal/Sheppard AFB	USA	0.87	2	251,653	240,460	0.96
LNK	Lincoln	USA	0.87	2	646,048	631,570	0.98
FSD	Sioux Falls	USA	0.87	2	2,088,917	2,044,050	0.98
BTV	Burlington (US) VT	USA	0.88	2	2,967,877	2,903,610	0.98
STC	St Cloud	USA	0.88	2	45,388	46,450	1.02
CMH	Columbus Port Columbus Intl Apt	USA	0.88	2	14,229,550	13,938,120	0.98
HOB	Hobbs Lea County Regional Apt	USA	0.88	2	76,642	70,050	0.91
SEA	Seattle-Tacoma International Apt	USA	0.88	2	52,926,494	51,924,550	0.98
SUN	Sun Valley Friedman Memorial Apt	USA	0.88	2	285,693	281,110	0.98
TXK	Texarkana	USA	0.88	2	177,744	173,620	0.98
BLV	Belleville	USA	0.88	2	37,010	36,710	0.99
CWA	Wausau Central Wisconsin Apt	USA	0.88	2	632,775	622,150	0.98
IDA	Idaho Falls	USA	0.88	2	740,021	727,960	0.98
SWF	Newburgh	USA	0.88	2	934,280	919,820	0.98
AGS	Augusta Bush Field	USA	0.88	2	1,311,708	1,292,590	0.99
CLL	College Station	USA	0.88	2	384,482	380,300	0.99
VPS	Destin-Ft Walton Beach Apt	USA	0.88	2	1,758,647	1,727,280	0.98
CNY	Moab	USA	0.88	2	25,145	22,450	0.89
PIH	Pocatello	USA	0.88	2	116,992	117,650	1.01
GSP	Greenville/Spartanburg Apt	USA	0.88	2	3,955,568	3,920,010	0.99

(Table 4.3 continues on next page)

(Table 4.3 continued)

Airport code	Airport name	Country	Random slope	Quantile group	Bookings	Passengers	Ratio
MFR	Medford	USA	0.88	2	1,415,879	1,397,430	0.99
OMA	Omaha Eppley Airfield	USA	0.88	2	9,348,981	9,239,820	0.99
ACT	Waco Regional Apt	USA	0.88	2	325,570	319,860	0.98
FNT	Flint	USA	0.89	2	2,059,843	2,033,510	0.99
BIL	Billings	USA	0.89	2	1,856,760	1,839,830	0.99
BGR	Bangor	USA	0.89	2	1,139,288	1,136,300	1.00
SJU	San Juan Luis Munoz Marin Intl Apt	Puerto Rico	0.89	2	15,582,750	15,422,880	0.99
CAK	Akron/Canton Ohio Regional	USA	0.89	2	3,974,304	3,945,820	0.99
AVP	Wilkes-Barre Scranton International Apt	USA	0.89	2	1,100,693	1,090,830	0.99
ALB	Albany International Airport	USA	0.89	2	5,833,349	5,794,160	0.99
LAS	Las Vegas McCarran International Apt	USA	0.89	2	74,569,285	74,001,180	0.99
SLK	Saranac Lake	USA	0.89	2	16,224	11,370	0.70
ASE	Aspen	USA	0.89	2	1,083,573	1,074,190	0.99
MKE	Milwaukee General Mitchell Intl Apt	USA	0.89	2	15,520,339	15,438,390	0.99
LBB	Lubbock Preston Smith International Apt	USA	0.89	2	2,328,279	2,315,360	0.99
AVL	Asheville	USA	0.89	2	1,725,598	1,722,650	1.00
SJT	San Angelo	USA	0.89	2	316,023	310,610	0.98
MHT	Manchester (US)	USA	0.89	2	6,130,663	6,125,020	1.00
ISP	Long Island Macarthur	USA	0.89	2	3,611,260	3,608,050	1.00
GJT	Grand Junction	USA	0.89	2	1,048,359	1,042,680	0.99
PFN	Panama City Bay County Apt	USA	0.89	2	22,517	21,920	0.97
RDU	Raleigh/Durham	USA	0.89	2	19,597,642	19,549,050	1.00
PSC	Pasco	USA	0.89	2	1,507,652	1,502,780	1.00
TEX	Telluride	USA	0.89	2	33,993	32,310	0.95
FAR	Fargo	USA	0.89	2	1,792,418	1,788,930	1.00

(Table 4.3 continues on next page)

(Table 4.3 continued)

Airport code	Airport name	Country	Random slope	Quantile group	Bookings	Passengers	Ratio
ORD	Chicago O'Hare International Apt	USA	0.89	2	66,609,586	66,588,890	1.00
GNV	Gainesville	USA	0.90	2	871,901	873,890	1.00
OME	Nome	USA	0.90	2	157,883	157,100	1.00
ATL	Atlanta Hartsfield-jackson Intl Apt	USA	0.90	2	62,515,596	62,734,610	1.00
BLI	Bellingham	USA	0.90	2	2,409,066	2,418,700	1.00
SLC	Salt Lake City	USA	0.90	2	24,165,486	24,247,150	1.00
PWM	Portland (US) ME	USA	0.90	2	3,990,080	4,033,560	1.01
SAT	San Antonio International Apt	USA	0.90	2	17,447,412	17,534,300	1.00
OAK	Oakland International Apt	USA	0.90	2	21,778,444	21,885,730	1.00
CLE	Cleveland Hopkins International Apt	USA	0.90	2	13,573,919	13,659,980	1.01
YKM	Yakima Air Terminal	USA	0.90	2	250,583	252,760	1.01
MBS	Saginaw	USA	0.90	2	617,675	623,070	1.01
MCO	Orlando International Apt	USA	0.90	2	66,735,463	67,270,530	1.01
BOI	Boise	USA	0.90	2	6,178,562	6,228,440	1.01
DCA	Washington Ronald Reagan National Apt	USA	0.90	2	32,900,624	32,931,260	1.00
TYR	Tyler	USA	0.90	2	400,981	404,440	1.01
SMX	Santa Maria (US)	USA	0.90	2	220,973	223,070	1.01
PIB	Laurel Hattiesburg-Laurel Regional Apt	USA	0.90	2	49,672	47,650	0.96
DHN	Dothan	USA	0.90	2	233,340	236,220	1.01
MIA	Miami International Apt	USA	0.90	2	23,604,776	23,852,020	1.01
BRW	Barrow	USA	0.90	3	158,729	158,810	1.00
TWF	Twin Falls	USA	0.90	3	147,362	148,550	1.01
DRO	Durango La Plata County Apt	USA	0.90	3	896,367	906,830	1.01
BET	Bethel Apt	USA	0.90	3	243,292	244,870	1.01
DTW	Detroit Wayne County	USA	0.91	3	33,319,495	33,738,120	1.01

(Table 4.3 continues on next page)

(Table 4.3 continued)

Airport code	Airport name	Country	Random slope	Quantile group	Bookings	Passengers	Ratio
GLH	Greenville (US) MS	USA	0.91	3	21,753	21,500	0.99
PNS	Pensacola International	USA	0.91	3	3,326,927	3,380,750	1.02
TPA	Tampa International Apt	USA	0.91	3	35,093,562	35,612,400	1.01
JAX	Jacksonville International Apt	USA	0.91	3	11,958,229	12,148,810	1.02
PSE	Ponce	Puerto Rico	0.91	3	462,957	469,250	1.01
GEG	Spokane International Apt	USA	0.91	3	6,758,459	6,875,550	1.02
BDL	Hartford Bradley International Apt	USA	0.91	3	12,763,773	12,979,180	1.02
BIS	Bismarck	USA	0.91	3	1,061,582	1,080,700	1.02
BOS	Boston Logan International Apt	USA	0.91	3	54,440,570	55,421,330	1.02
HGR	Hagerstown	USA	0.91	3	54,548	49,150	0.90
LAX	Los Angeles International Apt	USA	0.91	3	81,264,569	82,634,970	1.02
GTF	Great Falls International Apt	USA	0.91	3	836,865	852,710	1.02
ACY	Atlantic City International	USA	0.91	3	2,683,249	2,722,200	1.01
EKO	Elko	USA	0.91	3	103,636	106,040	1.02
DFW	Dallas/Fort Worth International Apt	USA	0.91	3	51,966,894	53,031,060	1.02
PVD	Providence	USA	0.91	3	8,938,442	9,146,930	1.02
HRL	Harlingen	USA	0.91	3	1,589,324	1,625,800	1.02
RUT	Rutland	USA	0.91	3	16,396	14,540	0.89
SJC	San Jose Norman Y. Mineta Intl	USA	0.92	3	19,192,098	19,664,910	1.02
PDX	Portland (US) OR	USA	0.92	3	26,931,037	27,616,270	1.03
DLH	Duluth	USA	0.92	3	721,471	740,630	1.03
AUS	Austin-Bergstrom International Apt	USA	0.92	3	20,465,694	21,025,670	1.03
BQN	Aguadilla	Puerto Rico	0.92	3	1,046,203	1,069,730	1.02
SAV	Savannah/Hilton Head International Apt	USA	0.92	3	3,709,252	3,820,460	1.03
EWR	Newark Liberty International Apt	USA	0.92	3	39,375,327	40,336,720	1.02

(Table 4.3 continues on next page)

(Table 4.3 continued)

Airport code	Airport name	Country	Random slope	Quantile group	Bookings	Passengers	Ratio
PSP	Palm Springs International Apt	USA	0.92	3	3,286,803	3,422,720	1.04
GFK	Grand Forks	USA	0.92	3	623,547	642,310	1.03
BZN	Bozeman	USA	0.92	3	1,960,551	2,020,650	1.03
HNL	Honolulu	USA	0.92	3	24,817,996	25,556,570	1.03
SMF	Sacramento International Apt	USA	0.92	3	19,640,268	20,271,650	1.03
MAF	Midland International Apt	USA	0.92	3	2,443,178	2,517,780	1.03
CPR	Casper	USA	0.92	3	443,310	455,970	1.03
EAT	Wenatchee	USA	0.92	3	240,917	249,150	1.03
GUC	Gunnison	USA	0.92	3	177,604	184,510	1.04
LGB	Long Beach	USA	0.92	3	6,580,131	6,796,890	1.03
LCH	Lake Charles	USA	0.92	3	313,266	323,520	1.03
EWN	New Bern	USA	0.92	3	627,949	651,030	1.04
IPT	Williamsport	USA	0.92	3	139,894	143,230	1.02
FAT	Fresno Yosemite International Airport	USA	0.92	3	2,648,587	2,740,340	1.03
STT	St Thomas Cyril E King Apt	Virgin Islands, US	0.93	3	2,758,053	2,848,840	1.03
DUJ	Dubois	USA	0.93	3	27,464	28,120	1.02
UTM	Tunica	USA	0.93	3	16,385	16,700	1.02
RNO	Reno	USA	0.93	3	7,720,620	8,002,930	1.04
ILG	Wilmington Greater Wilmington Apt	USA	0.93	3	157,392	162,650	1.03
HLN	Helena	USA	0.93	3	457,768	474,920	1.04
UIN	Quincy	USA	0.93	3	30,957	24,020	0.78
MSY	New Orleans Louis Armstrong Intl Apt	USA	0.93	3	19,354,625	20,125,400	1.04
MOT	Minot International Apt	USA	0.93	3	877,372	915,750	1.04
SDY	Sidney	USA	0.93	3	20,084	13,990	0.70
TTN	Philadelphia Trenton-Mercer Apt	USA	0.93	3	591,435	613,170	1.04

(Table 4.3 continues on next page)

(Table 4.3 continued)

Airport code	Airport name	Country	Random slope	Quantile group	Bookings	Passengers	Ratio
MSO	Missoula	USA	0.93	3	1,371,353	1,434,030	1.05
PBI	West Palm Beach International Apt	USA	0.93	3	13,164,236	13,757,790	1.05
RDM	Redmond/Bend	USA	0.93	3	1,053,541	1,096,050	1.04
BRO	Brownsville	USA	0.94	3	408,982	427,260	1.04
LYH	Lynchburg	USA	0.94	3	407,814	426,630	1.05
PHX	Phoenix Sky Harbor Intl Apt	USA	0.94	3	50,789,575	53,265,630	1.05
MFE	McAllen/Mission	USA	0.94	3	1,555,787	1,628,910	1.05
CKB	Clarksburg	USA	0.94	3	50,229	51,320	1.02
SBA	Santa Barbara	USA	0.94	3	1,692,881	1,775,380	1.05
PHL	Philadelphia International Apt	USA	0.94	3	36,608,206	38,391,140	1.05
LWS	Lewiston	USA	0.94	3	284,143	298,140	1.05
SGU	St George	USA	0.94	3	245,250	262,980	1.07
TUS	Tucson International Apt	USA	0.94	3	7,508,534	7,895,510	1.05
MEI	Meridian	USA	0.94	3	60,907	65,190	1.07
RSW	Fort Myers Sw Florida International Apt	USA	0.94	3	16,972,922	17,980,010	1.06
ABY	Albany Dougherty County Apt	USA	0.94	3	174,395	183,740	1.05
FLL	Fort Lauderdale/Hollywood Intl Apt	USA	0.94	3	41,904,841	44,194,300	1.05
ACV	Arcata/Eureka	USA	0.94	3	320,974	338,130	1.05
SFO	San Francisco	USA	0.94	3	58,784,697	62,028,940	1.06
TVC	Traverse City	USA	0.95	3	806,682	853,370	1.06
MYR	Myrtle Beach AFB	USA	0.95	3	3,506,727	3,723,290	1.06
PGV	Greenville (US) NC	USA	0.95	3	314,717	333,810	1.06
LIH	Lihue	USA	0.95	3	5,635,997	5,967,870	1.06
ABI	Abilene Regional Apt	USA	0.95	3	389,094	413,160	1.06
BPT	Beaumont/Port Arthur J. Brooks Regional	USA	0.95	3	118,352	129,090	1.09

(Table 4.3 continues on next page)

(Table 4.3 continued)

Airport code	Airport name	Country	Random slope	Quantile group	Bookings	Passengers	Ratio
ILM	Wilmington (US) NC	USA	0.95	3	1,765,308	1,884,320	1.07
SAN	San Diego International	USA	0.95	3	36,417,937	38,819,640	1.07
CHS	Charleston (US) SC	USA	0.95	3	5,813,660	6,230,530	1.07
MTJ	Montrose	USA	0.95	3	423,930	451,430	1.06
PUB	Pueblo	USA	0.95	3	49,455	44,130	0.89
JAC	Jackson (US) WY	USA	0.95	3	1,289,233	1,364,420	1.06
GTR	Columbus Golden Triangle Regional Apt	USA	0.95	3	180,585	193,140	1.07
PUW	Pullman/MOscow ID	USA	0.96	3	180,569	193,460	1.07
BUR	Burbank	USA	0.96	3	9,626,949	10,299,130	1.07
SRQ	Sarasota/Bradenton	USA	0.96	3	2,784,401	2,980,050	1.07
DEN	Denver Intl Apt	USA	0.96	3	61,679,658	66,134,280	1.07
ONT	Ontario	USA	0.96	3	9,763,343	10,510,230	1.08
AOO	Altoona	USA	0.96	3	20,448	21,570	1.05
MQT	Marquette	USA	0.97	3	219,438	233,280	1.06
UST	St Augustine	USA	0.97	3	23,861	25,490	1.07
SNA	Santa Ana	USA	0.97	3	19,546,474	21,211,330	1.09
PSM	Portsmouth Pease International Airport	USA	0.97	3	23,987	25,660	1.07
ALW	Walla Walla	USA	0.97	3	153,307	167,100	1.09
BFL	Bakersfield	USA	0.97	3	628,572	680,120	1.08
MGW	Morgantown	USA	0.97	3	42,031	45,960	1.09
ALO	Waterloo	USA	0.97	3	102,545	111,450	1.09
MRY	Monterey/Carmel Monterey Regional	USA	0.97	3	878,008	955,200	1.09
SHD	Staunton	USA	0.97	3	64,077	70,690	1.10
HYA	Hyannis	USA	0.97	3	33,147	17,610	0.53
SUX	Sioux City	USA	0.98	3	131,584	143,460	1.09

(Table 4.3 continues on next page)

(Table 4.3 continued)

Airport code	Airport name	Country	Random slope	Quantile group	Bookings	Passengers	Ratio
FCA	Kalispell	USA	0.98	3	862,353	945,020	1.10
JST	Johnstown	USA	0.98	3	34,708	37,840	1.09
EYW	Key West International Apt	USA	0.98	3	1,530,997	1,692,460	1.11
FLG	Grand Canyon Flagstaff Pulliam	USA	0.98	3	325,773	357,700	1.10
YUM	Yuma International Apt	USA	0.98	3	397,650	431,940	1.09
ITO	Hilo	USA	0.99	3	2,776,276	3,062,930	1.10
BTM	Butte	USA	0.99	3	124,067	139,320	1.12
GGG	Longview	USA	0.99	3	108,993	118,590	1.09
COD	Cody	USA	0.99	3	144,660	158,930	1.10
VLD	Valdosta Regional Apt	USA	0.99	3	183,958	204,750	1.11
FBS	Friday Harbor SPB	USA	0.99	3	18	20	1.11
ESC	Escanaba	USA	1.00	3	68,881	76,080	1.10
CMI	Champaign	USA	1.00	3	365,948	409,330	1.12
IMT	Iron Mountain	USA	1.00	3	47,576	52,430	1.10
COU	Columbia	USA	1.00	3	198,486	221,970	1.12
PRC	Prescott	USA	1.00	3	7,764	6,250	0.80
GCC	Gillette	USA	1.00	3	153,804	173,230	1.13
RKS	Rock Springs	USA	1.00	3	130,542	148,470	1.14
BJI	Bemidji	USA	1.00	3	106,362	119,610	1.12
ABR	Aberdeen (US)	USA	1.00	3	114,322	129,390	1.13
PPG	Pago Pago	American Samoa	1.00	3	105,440	118,260	1.12
DBQ	Dubuque	USA	1.01	3	157,460	177,900	1.13
STS	Santa Rosa (US)	USA	1.01	3	445,968	501,940	1.13
MAZ	Mayaguez	Puerto Rico	1.01	3	10,711	10,310	0.96
MLB	Melbourne	USA	1.01	3	917,701	1,040,590	1.13

(Table 4.3 continues on next page)

(Table 4.3 continues)

Airport code	Airport name	Country	Random slope	Quantile group	Bookings	Passengers	Ratio
DAB	Daytona Beach	USA	1.02	3	1,218,729	1,387,950	1.14
ALS	Alamosa	USA	1.02	3	23,540	24,710	1.05
PDT	Pendleton	USA	1.02	3	2,368	590	0.25
SPI	Springfield (US) IL	USA	1.03	3	333,176	378,300	1.14
CLT	Charlotte	USA	1.04	4	20,619,805	23,902,520	1.16
LBE	Latrobe	USA	1.04	4	333,877	387,040	1.16
IAH	Houston George Bush Intercont.	USA	1.04	4	26,410,615	30,471,770	1.15
BFD	Bradford	USA	1.04	4	11,603	13,130	1.13
IYK	Inyokern	USA	1.05	4	29,685	33,390	1.12
INL	International Falls	USA	1.05	4	65,071	77,600	1.19
LGA_int	New York La Guardia Apt	USA	1.05	4	6,455,199	7,660,330	1.19
MCW	Mason City	USA	1.05	4	33,426	34,130	1.02
PQI	Presque Isle	USA	1.05	4	55,468	66,520	1.20
BQK	Brunswick Glynco Jetport	USA	1.07	4	145,718	174,280	1.20
JHW	Jamestown (US) NY	USA	1.07	4	16,060	18,730	1.17
DVL	Devils Lake	USA	1.08	4	19,751	21,430	1.09
IRK	Kirksville	USA	1.08	4	15,815	10,940	0.69
LAR	Laramie	USA	1.09	4	43,839	55,470	1.27
RDD	Redding	USA	1.09	4	155,199	189,370	1.22
MSL	Muscle Shoals	USA	1.09	4	25,951	24,660	0.95
JMS	Jamestown (US) ND	USA	1.09	4	24,405	22,900	0.94
LNS	Lancaster	USA	1.10	4	12,064	9,290	0.77
CDC	Cedar City	USA	1.10	4	43,554	55,650	1.28
MMH	Mammoth Lakes	USA	1.10	4	108,745	128,620	1.18
TVF	Thief River Falls	USA	1.12	4	8,241	9,230	1.12

(Table 4.3 continues on next page)

(Table 4.3 continued)

Airport code	Airport name	Country	Random slope	Quantile group	Bookings	Passengers	Ratio
MSE	Manston	United Kingdom	1.12	4	38,754	48,185	1.24
CIU	Sault Ste Marie Chippewa County Apt	USA	1.12	4	86,658	108,450	1.25
EUG	Eugene	USA	1.12	4	1,681,856	1,843,300	1.10
MCE	Merced Regional Apt	USA	1.12	4	10,332	4,660	0.45
OTH	North Bend	USA	1.12	4	79,850	102,240	1.28
VIS	Visalia	USA	1.13	4	10,990	7,280	0.66
SBP	San Luis Obispo	USA	1.13	4	615,466	772,670	1.26
APN	Alpena	USA	1.14	4	50,267	64,620	1.29
PLN	Pellston	USA	1.14	4	107,649	136,960	1.27
ROW	Roswell	USA	1.14	4	154,449	196,920	1.27
BRD	Brainerd	USA	1.15	4	70,737	91,940	1.30
ATY	Watertown (US) SD	USA	1.16	4	25,898	28,930	1.12
CLD	San Diego McClellan-Palomar Arpt	USA	1.17	4	207,451	268,780	1.30
HIB	Hibbing/Chisholm	USA	1.18	4	50,520	65,800	1.30
SOW	Show Low	USA	1.18	4	8,968	6,430	0.72
MLS	Miles City	USA	1.18	4	685	1,050	1.53
IWD	Ironwood	USA	1.20	4	5,994	4,970	0.83
IPL	Imperial County Apt	USA	1.20	4	18,136	25,130	1.39
DRT	Del Rio International Apt	USA	1.22	4	31,473	39,100	1.24
HON	Huron	USA	1.23	4	5,078	4,650	0.92
EAU	Eau Claire	USA	1.23	4	97,187	134,680	1.39
PAH	Paducah	USA	1.24	4	96,071	133,890	1.39
LMT	Klamath Falls	USA	1.24	4	65,107	89,030	1.37
MKG	Muskegon	USA	1.25	4	74,234	104,480	1.41
FOD	Fort Dodge	USA	1.26	4	24,995	29,070	1.16

(Table 4.3 continues on next page)

(Table 4.3 continued)

Airport code	Airport name	Country	Random slope	Quantile group	Bookings	Passengers	Ratio
MOD	Modesto	USA	1.26	4	66,712	96,470	1.45
MCK	Mccook	USA	1.30	4	5,880	7,300	1.24
CMX	Hancock	USA	1.30	4	109,656	159,820	1.46
CIC	Chico	USA	1.33	4	80,907	122,460	1.51
CEC	Crescent City	USA	1.33	4	56,308	84,100	1.49
VCT	Victoria	USA	1.34	4	15,802	21,490	1.36
JBR	Jonesboro	USA	1.34	4	518	680	1.31
TBN	Fort Leonard Wood	USA	1.35	4	31,627	42,000	1.33
BRL	Burlington (US) IA	USA	1.45	4	1,104	1,540	1.39
FKL	Franklin Chess Lamberton Apt	USA	1.46	4	4,773	7,170	1.50
WRL	Worland	USA	1.47	4	7,141	9,520	1.33
GBD	Great Bend	USA	1.47	4	2,716	3,090	1.14
MCN	Macon Lewis B Wilson Apt	USA	1.48	4	5,537	4,650	0.84
ANI	Aniak	USA	1.51	4	2,691	360	0.13
CVN	Clovis Municipal Apt	USA	1.52	4	4,161	2,960	0.71
LDY	Derry	United Kingdom	1.54	4	119,151	221,932	1.86
CDR	Chadron	USA	1.55	4	4,827	7,960	1.65
SXP	Nunam Iqua	USA	1.55	4	10	20	2.00
JER	Jersey	United Kingdom	1.67	4	174,385	336,860	1.93
SWF_dom	Newburgh	USA	1.68	4	1,002,016	1,886,366	1.88
STG	St George Island	USA	1.68	4	9	20	2.22
IGM	Kingman	USA	1.69	4	2,701	2,220	0.82
PIK	Glasgow Prestwick Apt	United Kingdom	1.73	4	2,689,574	5,177,870	1.93
SEN	London Southend Apt	United Kingdom	1.75	4	1,137,845	2,271,988	2.00
LPL	Liverpool	United Kingdom	1.75	4	9,470,665	18,597,862	1.96

(Table 4.3 continues on next page)

(Table 4.3 continued)

Airport code	Airport name	Country	Random slope	Quantile group	Bookings	Passengers	Ratio
LTN	London Luton Apt	United Kingdom	1.84	4	21,753,310	45,183,148	2.08
SPB	St Thomas Charlotte Amalie SPB	Virgin Islands, US	1.84	4	8	20	2.50
AIA	Alliance	USA	1.84	4	3,629	6,340	1.75
STN	London Stansted Apt	United Kingdom	1.87	4	41,840,667	87,957,362	2.10
CPX	Culebra	Puerto Rico	1.88	4	605	1,040	1.72
LBA	Leeds Bradford	United Kingdom	1.91	4	6,173,042	13,364,266	2.16
BOH	Bournemouth	United Kingdom	1.91	4	1,564,152	3,274,628	2.09
GCI	Guernsey	United Kingdom	1.93	4	32,416	54,948	1.70
EMA	Nottingham East Midlands Airport	United Kingdom	1.94	4	8,653,193	19,138,478	2.21
LGA_dom	New York La Guardia Apt	USA	1.96	4	58,714,031	129,161,074	2.20
ELY	Ely	USA	1.98	4	690	240	0.35
DSA	Doncaster/Sheffield	United Kingdom	2.01	4	1,560,236	3,476,071	2.23
BRS	Bristol (GB) 00	United Kingdom	2.02	4	10,889,003	24,756,936	2.27
LCY	London City Apt	United Kingdom	2.03	4	5,817,034	13,246,107	2.28
JFK_int	New York J F Kennedy International Apt	USA	2.05	4	59,017,777	136,143,142	2.31
GLO	Gloucester/Cheltenham	United Kingdom	2.06	4	166	441	2.66
BLK	Blackpool	United Kingdom	2.08	4	372,326	886,221	2.38
BFS	Belfast International Apt	United Kingdom	2.09	4	2,943,541	7,058,851	2.40
EDI	Edinburgh	United Kingdom	2.16	4	10,307,629	25,058,001	2.43
GGW	Glasgow	USA	2.20	4	1,334	2,250	1.69
HVR	Havre	USA	2.20	4	1,164	1,880	1.62
KOI	Kirkwall	United Kingdom	2.22	4	347	435	1.25
SVC	Silver City	USA	2.22	4	2,992	1,370	0.46
EXT	Exeter (GB) 00	United Kingdom	2.26	4	827,126	2,078,031	2.51
SOU	Southampton	United Kingdom	2.28	4	1,314,100	3,389,175	2.58

(Table 4.3 continues on next page)

(Table 4.3 continued)

Airport code	Airport name	Country	Random slope	Quantile group	Bookings	Passengers	Ratio
LGW	London Gatwick Apt	United Kingdom	2.28	4	60,134,571	155,048,677	2.58
OGS	Ogdensburg	USA	2.38	4	4,299	5,810	1.35
INV	Inverness	United Kingdom	2.41	4	56,916	152,700	2.68
EWR_int	Newark Liberty International Apt	USA	2.41	4	22,319,934	60,597,950	2.71
LWT	Lewistown	USA	2.48	4	293	460	1.57
JFK_dom	New York J F Kennedy International Apt	USA	2.50	4	45,951,975	128,904,173	2.81
BHX	Birmingham Airport	United Kingdom	2.52	4	14,103,311	39,797,043	2.82
CNM	Carlsbad	USA	2.54	4	2,122	2,370	1.12
MAN	Manchester (GB)	United Kingdom	2.60	4	31,007,267	90,446,148	2.92
NCL	Newcastle	United Kingdom	2.63	4	5,548,510	16,130,750	2.91
OLF	Wolf Point	USA	2.64	4	2,600	2,290	0.88
EWR_dom	Newark Liberty International Apt	USA	2.66	4	41,102,901	122,359,098	2.98
GAL	Galena	USA	2.76	4	353	80	0.23
MSS	Massena	USA	2.84	4	5,286	9,420	1.78
CWL	Cardiff (GB) 00	United Kingdom	2.87	4	1,553,931	4,561,283	2.94
GLA	Glasgow International Airport	United Kingdom	2.90	4	5,684,941	18,553,184	3.26
LHR	London Heathrow Apt	United Kingdom	2.98	4	101,883,653	341,385,643	3.35
LAM	Los Alamos	USA	3.04	4	24	190	7.92
ABZ	Aberdeen (GB)	United Kingdom	3.14	4	1,333,790	4,721,406	3.54
SSB	St Croix SPB	Virgin Islands, US	3.18	4	31	80	2.58
AHN	Athens (US)	USA	3.29	4	291	680	2.34
BID	Block Island	USA	3.32	4	8	40	5.00
MME	Durham	United Kingdom	3.37	4	193,800	678,364	3.50
IFP	Bullhead City	USA	3.47	4	38,154	139,330	3.65
GDV	Glendive	USA	3.48	4	759	1,840	2.42

(Table 4.3 continues on next page)

(Table 4.3 continued)

Airport code	Airport name	Country	Random slope	Quantile group	Bookings	Passengers	Ratio
IOM	Isle of Man	United Kingdom	3.49	4	55,967	196,496	3.51
NWI	Norwich	United Kingdom	3.70	4	258,319	1,032,095	4.00
BHD	Belfast George Best City Apt	United Kingdom	3.93	4	209,537	487,209	2.33
HUY	Humberside	United Kingdom	4.71	4	152,113	753,649	4.95
ALM	Alamogordo Municipal Apt	USA	4.91	4	82	280	3.41
MKL	Jackson	USA	4.98	4	195	270	1.38
HNM	Hana	USA	5.25	4	9	40	4.44
SCM	Scammon Bay	USA	5.30	4	2	20	10.00
NQY	Newquay	United Kingdom	7.45	4	29,261	44,935	1.54
MLL	Marshall	USA	9.01	4	1	20	20.00
WIC	Wick	United Kingdom	9.98	4	38	398	10.47
EWB	New Bedford	USA	13.54	4	3	60	20.00
LSI	Shetland Islands Sumburgh Apt	United Kingdom	24.86	4	3,277	13,750	4.20
USA	Concord	USA	2983.51	4	3	18,130	6043.33

Table 4.4: Random slope coefficient values for each country present in TravelPac and OAG data, with their respective number of passengers and bookings and the ratios calculated as passengers per bookings.
Note: the quartile groups 1 to 4 are those shown in **Figure 4.7**.

Country name	Random slope	Quantile group	Bookings	Passengers	Ratio
Turkmenistan	0.12	1	3,045	600.47	0.20
Falkland Islands	0.25	1	2,044	537.36	0.26
Norway	0.32	1	81,412	43,737.87	0.54
Singapore	0.37	1	37,279	23,634.15	0.63
Barbados	0.37	1	24,175	17,980.49	0.74
Hong Kong (sar) China	0.39	1	52,515	36,156.41	0.69
Costa Rica	0.41	1	3,357	2,819.27	0.84
Denmark	0.42	1	71,122	41,052.33	0.58
Russian Federation	0.42	1	23,635	12,221.13	0.52
Algeria	0.43	1	8,310	2,725.80	0.33
Uruguay	0.44	1	345	77.39	0.22
Sweden	0.45	1	63,083	62,879.03	1.00
Angola	0.45	1	1,779	2,609.58	1.47
Switzerland	0.45	1	272,540	304,337.23	1.12
Fiji	0.48	1	820	979.97	1.20
Bahrain	0.48	1	7,889	6,028.59	0.76
Bermuda	0.49	1	4,008	431.53	0.11
Azerbaijan	0.49	1	2,980	2,531.39	0.85
Israel	0.49	1	17,364	19,430.99	1.12
Indonesia	0.49	1	5,692	2,933.86	0.52
Luxembourg	0.50	1	10,235	13,410.23	1.31
Faroe Islands	0.50	1	572	340.50	0.60
Iceland	0.51	1	16,523	17,709.41	1.07
Ireland Republic of	0.51	1	360,516	292,546.30	0.81
Saint Lucia	0.51	1	9,659	5,054.37	0.52
Guatemala	0.52	1	699	924.05	1.32

(Table 4.4 continues on next page)

(Table 4.4 continued)

Country name	Random slope	Quantile group	Bookings	Passengers	Ratio
Kuwait	0.52	1	4,269	3,650.48	0.86
Bahamas	0.53	1	2,426	1,599.60	0.66
Jordan	0.53	1	5,401	4,468.00	0.83
Oman	0.54	1	8,074	9,668.17	1.20
Slovakia	0.54	1	23,918	23,685.60	0.99
Germany	0.55	1	370,813	358,104.39	0.97
Nigeria	0.56	1	30,597	32,777.28	1.07
Brunei Darussalam	0.57	1	524	481.29	0.92
Saudi Arabia	0.59	1	14,605	15,430.71	1.06
Syrian Arab Republic	0.59	1	6,109	6,397.97	1.05
Korea Republic of	0.60	1	4,376	5,877.42	1.34
Sudan	0.60	1	2,014	3,167.38	1.57
Lebanon	0.60	1	5,374	2,621.49	0.49
Tanzania United Republic of	0.60	1	5,823	6,375.63	1.09
Puerto Rico	0.61	1	4,147	3,413.47	0.82
Georgia	0.63	1	1,050	1,273.76	1.21
Albania	0.63	1	3,340	9,373.23	2.81
Mongolia	0.63	1	834	714.07	0.86
Japan	0.64	2	13,055	22,532.57	1.73
Belgium	0.64	2	42,283	51,230.36	1.21
Mauritius	0.64	2	12,223	11,356.21	0.93
United Arab Emirates	0.64	2	102,106	133,335.04	1.31
USA	0.65	2	475,051	566,019.87	1.19
Latvia	0.65	2	15,811	14,160.26	0.90
Estonia	0.65	2	4,934	6,135.42	1.24

(Table 44 continues on next page)

(Table 4.4 continued)

Country name	Random slope	Quantile group	Bookings	Passengers	Ratio
Gibraltar	0.65	2	9,492	14,504.86	1.53
Maldives	0.66	2	15,528	16,263.75	1.05
Saint Kitts and Nevis, Leeward Islands	0.66	2	2,105	3,368.42	1.60
Libya	0.66	2	7,816	8,536.33	1.09
Ethiopia	0.66	2	3,566	5,271.89	1.48
Malaysia	0.67	2	27,646	30,294.13	1.10
Italy	0.68	2	296,363	381,569.13	1.29
Lithuania	0.68	2	18,838	18,078.53	0.96
Zambia	0.69	2	2,054	1,269.88	0.62
Thailand	0.69	2	72,382	117,542.09	1.62
Canada	0.69	2	47,435	68,786.01	1.45
Sri Lanka	0.71	2	18,005	23,674.39	1.31
South Africa	0.71	2	69,500	132,434.49	1.91
Romania	0.71	2	20,841	31,186.87	1.50
Cayman Islands	0.72	2	1,727	1,626.07	0.94
Iran Islamic Republic of	0.72	2	7,600	6,296.62	0.83
Peru	0.72	2	1,956	1,272.33	0.65
Kazakhstan	0.74	2	2,833	3,069.29	1.08
Hungary	0.74	2	32,073	52,854.40	1.65
Finland	0.74	2	19,228	12,036.16	0.63
Seychelles	0.74	2	1,879	457.66	0.24
Uganda	0.74	2	4,359	7,216.23	1.66
Belarus	0.75	2	1,331	872.80	0.66
Chile	0.75	2	2,318	3,530.42	1.52
Qatar	0.76	2	8,176	9,895.16	1.21

(Table 4.4 continues on next page)

(Table 4.4 continued)

Country name	Random slope	Quantile group	Bookings	Passengers	Ratio
Namibia	0.76	2	860	991.08	1.15
Armenia	0.77	2	416	702.76	1.69
Netherlands	0.78	2	174,269	274,514.24	1.58
Czech Republic	0.78	2	44,395	73,252.45	1.65
China	0.79	2	33,672	51,818.75	1.54
Egypt	0.80	2	72,302	183,194.40	2.53
India	0.80	2	138,584	276,666.33	2.00
Ukraine	0.80	2	6,509	9,329.28	1.43
Cambodia	0.81	2	1,646	541.97	0.33
Poland	0.82	2	156,728	228,409.62	1.46
Serbia	0.82	2	4,127	8,096.60	1.96
Brazil	0.82	2	17,063	32,529.77	1.91
Argentina	0.82	3	5,962	9,553.15	1.60
Belize	0.82	3	859	954.39	1.11
Dominica	0.83	3	800	953.26	1.19
Australia	0.83	3	75,384	161,429.76	2.14
Haiti	0.83	3	387	392.94	1.02
Antigua and Barbuda, Leeward Islands	0.84	3	8,623	10,025.59	1.16
Panama	0.84	3	1,130	1,318.05	1.17
Philippines	0.84	3	11,690	16,464.26	1.41
New Zealand	0.87	3	29,845	80,308.04	2.69
Croatia	0.87	3	7,148	7,175.72	1.00
Nepal	0.90	3	3,653	6,480.11	1.77
Ecuador	0.91	3	1,467	4,330.90	2.95
Ghana	0.91	3	8,008	16,255.95	2.03

(Table 4.4 continues on next page)

(Table 4.4 continued)

Country name	Random slope	Quantile group	Bookings	Passengers	Ratio
Trinidad and Tobago	0.94	3	5,719	13,415.46	2.35
Morocco	0.94	3	41,698	84,871.93	2.04
Gabon	0.99	3	672	856.96	1.28
Bolivia	1.00	3	793	471.12	0.59
Kyrgyzstan	1.01	3	354	482.70	1.36
Mali	1.02	3	277	424.22	1.53
Eritrea	1.02	3	367	679.62	1.85
Pakistan	1.02	3	38,877	110,727.27	2.85
Kenya	1.02	3	16,648	44,073.56	2.65
Turkey	1.02	3	52,506	79,352.63	1.51
Cyprus	1.03	3	49,894	75,493.52	1.51
Slovenia	1.03	3	4,352	10,274.22	2.36
Austria	1.05	3	59,093	162,048.30	2.74
Uzbekistan	1.08	3	990	422.58	0.43
Yemen	1.08	3	942	2,393.20	2.54
Portugal	1.09	3	115,195	193,193.80	1.68
Grenada, Windward Islands	1.09	3	3,254	9,111.78	2.80
France	1.11	3	271,331	755,972.37	2.79
Rwanda	1.13	3	559	1,085.19	1.94
Colombia	1.15	3	2,013	6,589.16	3.27
Spain	1.16	3	659,647	1,499,430.50	2.27
Congo	1.17	3	425	769.71	1.81
Bulgaria	1.18	3	18,308	51,380.73	2.81
Viet Nam	1.19	3	5,570	14,415.06	2.59
Venezuela	1.19	3	1,094	1,108.99	1.01

(Table 4.4 continues on next page)

(Table 4.4 continued)

Country name	Random slope	Quantile group	Bookings	Passengers	Ratio
Malta	1.20	3	28,636	52,631.91	1.84
Bangladesh	1.20	3	7,673	20,985.11	2.73
Greece	1.21	3	31,787	29,699.81	0.93
Mauritania	1.24	3	268	519.16	1.94
Madagascar	1.29	3	136	477.46	3.51
Benin	1.31	3	349	722.23	2.07
Jamaica	1.35	4	17,378	66,547.38	3.83
Virgin Islands, British	1.35	4	543	2,693.46	4.96
Myanmar	1.40	4	452	1,930.52	4.27
El Salvador	1.45	4	168	392.94	2.34
Botswana	1.46	4	444	1,700.07	3.83
Cameroon	1.47	4	720	3,626.78	5.04
Zimbabwe	1.48	4	2,772	8,243.12	2.97
Bosnia and Herzegovina	1.56	4	597	1,902.49	3.19
Nicaragua	1.57	4	369	948.85	2.57
Mozambique	1.60	4	272	1,147.10	4.22
Cote D'Ivoire	1.61	4	835	1,019.37	1.22
Senegal	1.63	4	489	1,161.70	2.38
Dominican Republic	1.63	4	11,056	9,464.87	0.86
Guyana	1.71	4	440	2,663.80	6.05
French Polynesia	1.73	4	221	638.61	2.89
Mexico	1.76	4	13,858	56,840.96	4.10
Turks and Caicos Islands	1.78	4	818	2,798.28	3.42
Malawi	1.88	4	484	1,392.78	2.88
Macedonia Former Yugoslav Republic of	1.97	4	543	3,252.49	5.99

(Table 4.4 continues on next page)

(Table 4.4 continued)

Country name	Random slope	Quantile group	Bookings	Passengers	Ratio
Sierra Leone	1.98	4	1,154	6,558.69	5.68
Iraq	1.99	4	837	861.38	1.03
St Vincent and the Grenadines	2.19	4	873	474.21	0.54
Djibouti	2.23	4	257	1,006.65	3.92
Montenegro	2.29	4	1,118	3,776.93	3.38
Cuba	2.30	4	10,762	40,196.86	3.74
Honduras	2.52	4	679	3,060.83	4.51
Liberia	2.52	4	143	702.76	4.91
Paraguay	2.52	4	162	732.02	4.52
Moldova Republic of	2.58	4	713	3,796.39	5.32
French Guiana	2.72	4	140	691.35	4.94
Swaziland	2.80	4	81	415.57	5.13
Tunisia	3.28	4	13,227	54,533.34	4.12
Togo	3.37	4	226	1,434.71	6.35
Afghanistan	3.95	4	187	582.84	3.12
Gambia	4.72	4	2,575	16,800.78	6.52
Niger	4.85	4	57	554.59	9.73
Lao People's Democratic Republic	5.27	4	91	1,465.63	16.11
Somalia	5.39	4	110	1,206.60	10.97
Burkina Faso	5.75	4	340	4,012.89	11.80
Guadeloupe	6.87	4	55	806.63	14.67
Monaco	10.74	4	121	3,035.79	25.09
Papua New Guinea	18.83	4	15	696.28	46.42
Macao (sar) China	45.22	4	13	1,592.48	122.50
Greenland	54.36	4	4	679.08	169.77

Table 6.2: Pearson correlation coefficient and confidence intervals for each factor used, to determine a single proxy for healthcare (**Chapter 6**).

	Health expenditure	HIV treatment	Life expectancy	Measles vaccination	Tuberculosis incidence
Health expenditure					
HIV treatment	0.27 (0.11, 0.40)				
Life expectancy	0.40 (0.28, 0.50)	0.34 (0.19, 0.47)			
Measles vaccination	0.24 (0.11, 0.35)	0.35 (0.20, 0.48)	0.63 (0.54, 0.70)		
Tuberculosis incidence	-0.12 (-0.25, 0.01)	-0.12 (0.28, 0.04)	-0.66 (-0.73, -0.57)	-0.35 (-0.46, -0.22)	