an introduction to

Modelling population flows using spatial interaction models

Adamon Dennett* University College London

* Corresponding author. Email: a.dennett@ucl.ac.uk. Address: Bartlett Centre for Advanced Spatial Analysis, University College London, Gower Street, London, WC1E 6BT, UK

Paper received 18 April 2018; accepted 10 May 2018; published 12 November 2018

Abstract

Background

Spatial Interaction Models have been used for decades to explain and predict flows (of migrants, capital, traffic, trade etc.) between geographic locations.

Aims

This paper will guide users through the process of fitting and calibrating spatial interaction models in order to understand, explain and predict internal migration flows in Australia.

Data and methods

Internal migration data from the Australian 2011 Census of Population and Housing, which records people who have moved between Greater Capital City Statistical Areas over 5-year periods, is used to exemplify the modelling process. The R statistical software is used to process and visualise the data as well as run the models.

Results

The full suite of Wilson’s family of spatial interaction models is fitted to the internal migration data, revealing that distance and origin/destination populations are some of the most important influencing factors affecting internal migration flows. We see whether constraining the model to known flows about origins and/or destinations will improve the fits and model estimates.

Conclusions

Spatial interaction modelling has been a tool in the box of some population geographers for a number of decades. However, recent advances in more forgiving programming languages such as R and Python now mean that this powerful modelling methodology is no longer only available to those who also possess advanced computer programming skills. This guide has exemplified the process of fitting and calibrating spatial interaction models on Australian internal migration data, but the methods could easily be applied to other flow data sets in other contexts.

Key words

Spatial Interaction Model; Gravity Model; Migration; R; Estimation; Visualisation; Census; Australia.

1. Introduction

In this introductory guide, you will learn how we can use spatial interaction models to model population flows for a variety of different purposes such as estimating unknown flows, predicting future patterns, understanding the drivers of those flows, or exploring the differences between the flows of different groups. The empirical example uses migration flows taken from the Australian Bureau of Statistics (ABS) 2011 Census of Population and Housing (Census), but the method is generic and could be used on any flow data (other population data such as commuting data or economic data such as flows of capital or trade, for example). The examples shown here will use the R software environment, and an accompanying practical walk-through guide will be referred to throughout this paper which can be accessed via the following link: https://rpubs.com/adam_dennett/376877. It can also be followed in its entirety if you would like to learn the code required for any of the models.

1.1 What constitutes a population flow?

In fields such as population geography and demography, the population flows of interest are usually low-frequency migration or residential mobility moves. Both assume some permanent change of residential address which can be either within a country (internal migration or residential mobility) or between countries (international migration). Some scholars make a clear distinction between what they term ‘residential mobility’ (short-distance moves, usually within settlements or regions where individuals may retain the same social groups or job) and ‘internal migration’ (longer distance moves, which may involve changing jobs and social groups).

However, in reality this is a continuum with no clear line demarking one or the other. Most national censuses will simply refer to any internal move over any distance as ‘internal migration’. Depending on where you are in the world, the proportion of a population changing their residential address in a given year is around 10 per cent, with most people only moving a handful of times in their lifetime. These population movements can be contrasted with high-frequency flows that occur on daily or weekly cycles. The commute to work or school or travel to the shops, while of interest to some population geographers, is frequently the domain of transport planners and analysts who are concerned with the impact of these moves on transport infrastructure and systems.

1.2. Population flow data

Data related to these flows can vary. For migration, data captured by censuses or surveys tend to be ‘transitions’ over a period of time (a year, 5 years, 10 years) with a flow recorded if there is a difference between residential address between the start and the end of this period (Rees 1977). Transition data will not record multiple moves during a period and, as such, are simpler than movement data which record multiple moves. Movement data can be found more commonly in population registers that track population changes continuously. Commuting data, while frequently transitions derived from census returns, are increasingly being obtained from sources such as mobile phone cell tower connections and apps (Erhardt and Dennett 2017). In this guide we will use census-based transition data; however, being aware of these other sources is important, particularly in a world that is looking increasingly to move away from censuses.

1.3. Why would we want to model population flows?

As hinted in the introduction, there are a number of reasons for wanting to model population flows. In migration studies, in a number of papers by Raymer and colleagues (Raymer 2007; Raymer, Abel and Smith 2007; Raymer and Abel 2008) as well Abel (2010), Willekens (1999) and Dennett and Wilson (2013), the problem of missing or incomplete data was addressed. Pooler (1987) used similar models to tackle the problem of predicting migration, as did Fotheringham et al. (2001). Implicit in much of this modelling research is the evaluation of factors influencing the observed patterns – a paper by Kim and Cohen (2010) is more explicit in this regard.

1.4. How can we model population flows?

Population flows can be conceptualised as interactions between two entities – origins and destinations – which have different properties of emissivity and attractiveness (see Lee 1966 for the classic paper on this topic in relation to migration). The strength of interaction is a function of these origin and destination properties and the negative influence of the cost (frequently some measure of distance, but equally could be financial, time or some other cost) that might be associated with travelling between them. This situation is analogous to that observed by Newton when he defined the laws of gravity – the larger the entities interacting, the stronger the force of interaction; the further the distance between them, the weaker the interaction. Hence the term 'gravity model' has been used and the equation applied in studying population flows for a long time. See Zipf (1946) for one of the earlier applications of the model to population flows.

Over the years, various improvements have been made to the basic gravity model. Perhaps the most important paper in this respect is by Wilson (1971), where he introduces the idea of 'constraints' which force the flows or interactions estimated by the model to adhere to known information about the system. For example, there may be data on the total number of people leaving an origin or arriving at a destination (or both). In the basic gravity model, the flows estimated might exceed this known information, which is clearly an issue. By introducing constraints, it is possible to force the modelled flows to correspond to this known information, significantly improving accuracy. Wilson called his new family of constrained models 'spatial interaction models'.

2. Modelling population flows in practice

The basic theory behind spatial interaction or gravity models is not too difficult to comprehend. Where the challenges begin is in running a spatial interaction model in practice. In the 1970s and 1980s, when spatial interaction modelling was being established as part of the tool kit for people working on spatial problems, running a model would require some fairly high-level computer programming knowledge. Today, however, software has advanced and more forgiving languages such as Python and R mean that spatial interaction models can be used as a tool by many more researchers.

In this guide we will be using R, with details of the full implementation referred to in this paper found in an accompanying walk-through guide designed to be worked through while reading the text here. The guide includes all of the data and code you need to run a spatial interaction model, and can be accessed via the link at the beginning of this paper[1]. If, however, you would like to explore these models in Python, then Oshan (2016) has written an excellent primer that is worth reading, while Dan Lewis has translated a similar R walk-through of mine into Python using UK data[2]. For consistency, Oshun's notation is adopted in this paper.

2.1. Data

To illustrate the modelling exercise, migration data (derived from the answer about previous residence 5 years ago – therefore comprising 5-year transitions) from the 2011 Census have been obtained. These data are at the Greater Capital City Statistical Area (GCCSA) level, which is comprised of 15 zones (Figure 1).

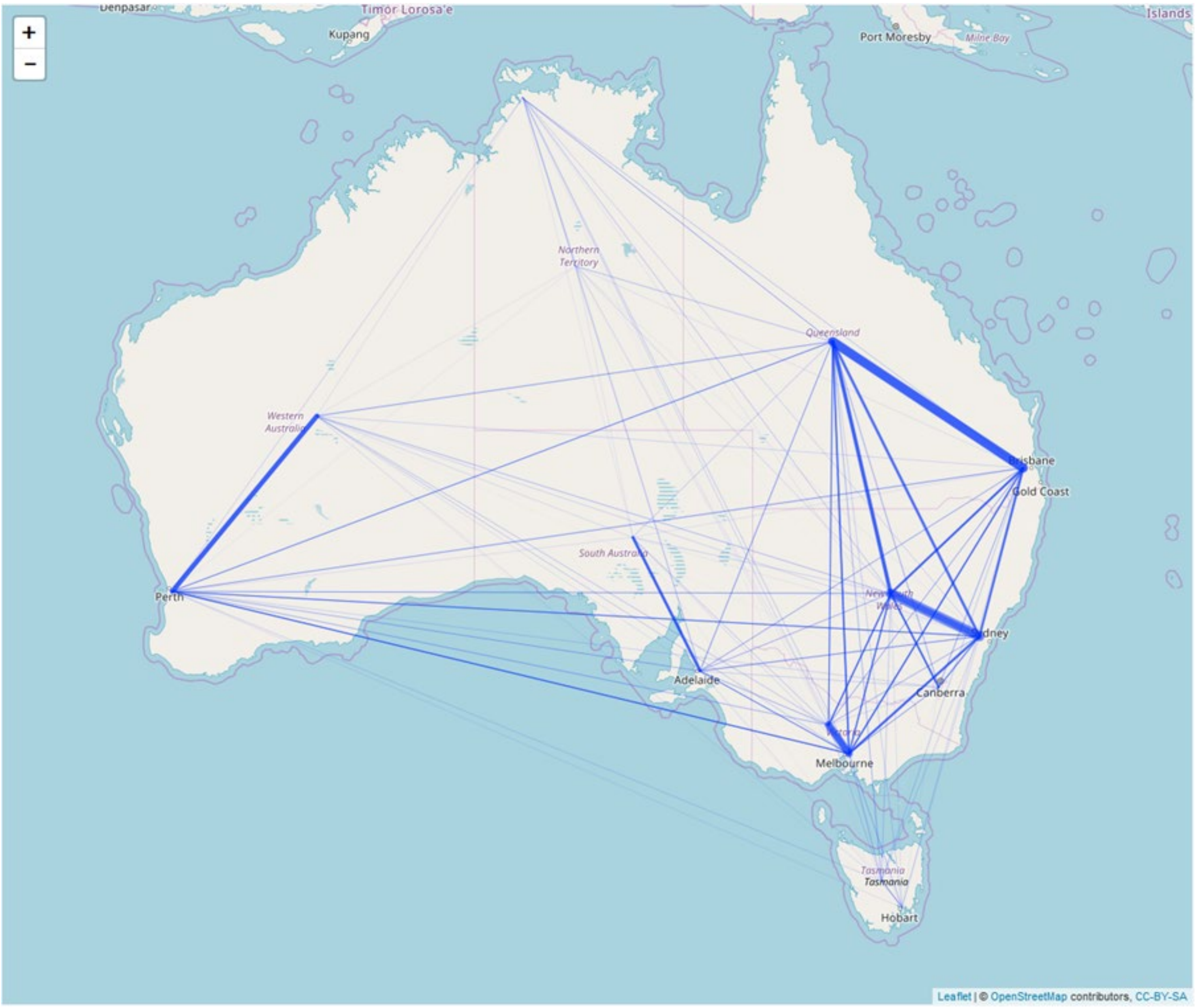


Figure 1: Five-year migration flows between GCCSAs, 2006–2011

Source: ABS 2011 Census. *Note*: Line weights indicative of volumes.

[1] https://rpubs.com/adam_dennett/376877

[2] https://github.com/danlewis85/UCL_CASA_Urban_Simulation

Accompanying the migration flows are variables for each origin and destination relating to:

- total population
- unemployment rate
- median weekly income
- percentage of households living in rented accommodation.

These variables can be used to try and explain observed migration flows or predict flows if none are available. A table containing origin/destination pairs, the flows between them and these origin and destination specific variables can be observed in Table 1 and is downloadable in the accompanying exercise.

Table 1: Sample of pair-wise migration flow data with accompanying data relating to origin and destination characteristics

Origin	Destination	Flow	OrigPop	DestPop	Orig Unemp	Dest Unemp	Orig Med Income	Dest Med Income	Orig % Rented	Dest % Rented
Greater Sydney	Greater Sydney	3,395,019	4,391,673	4,391,673	5.74	5.74	780.64	780.64	31.77	31.77
Greater Sydney	Rest of NSW	91,043	4,391,673	2,512,952	5.74	6.12	780.64	509.97	31.77	27.20
Greater Sydney	Greater Melbourne	22,605	4,391,673	3,999,981	5.74	5.47	780.64	407.95	31.77	27.34
Greater Sydney	Rest of Vic.	4,420	4,391,673	1,345,717	5.74	5.17	780.64	506.58	31.77	24.08
Greater Sydney	Greater Brisbane	22,874	4,391,673	2,065,998	5.74	5.86	780.64	767.08	31.77	33.19
Greater Sydney	Rest of Qld	27,447	4,391,673	2,253,723	5.74	6.22	780.64	446.48	31.77	32.57
Greater Sydney	Greater Adelaide	5,829	4,391,673	1,225,235	5.74	5.78	780.64	445.53	31.77	28.27
Greater Sydney	Rest of SA	795	4,391,673	368,260	5.74	5.45	780.64	522.71	31.77	26.17
Greater Sydney	Greater Perth	10,572	4,391,673	1,728,865	5.74	4.76	780.64	730.84	31.77	27.52

Source: ABS 2011 Census

2.2. The ‘unconstrained’ / ‘total constrained’ spatial interaction model

2.2.1. The multiplicative modelling framework

The classic gravity model, which estimates flows/interactions as a function of predictor variables (a model virtually identical to that used by Zipf), can be written as follows:

$$T_{ij} = k\frac{V_i^{\mu}W_j^{\alpha}}{d_{ij}^{\beta}} \qquad (1)$$

This model can be rearranged and written in the multiplicative form more familiar from Wilson’s 1971 paper:

$$T_{ij} = kV_i^{\mu}W_j^{\alpha}d_{ij}^{-}\beta \qquad (2)$$

where:

T_{ij} is the transition/trip or flow, T, between origin i (always the rows in a matrix) and destination j (always the columns in a matrix). If you are not overly familiar with matrix notation, the i and j are just generic indexes to allow us to refer to any cell in the matrix.

V is a vector (a 1 dimensional matrix – or, if you like, a single line of numbers) of origin attributes which relate to the emissivity of all origins in the dataset, i – this could be any of the origin-related variables in Table 1.

W is a vector of destination attributes relating to the attractiveness of all destinations in the dataset, j – similarly, this could be any of the destination related variables in Table 1.

d is a matrix of costs (frequently distances – hence, *d*) relating to the flows between i and j.

k, μ, α and β are all model parameters to be estimated. β is assumed to be negative, as with an increase in cost/distance we would expect interaction to decrease.

Wilson's term for this basic gravity model is the 'unconstrained' model. However, k is a constant of proportionality and forces all flow estimates to add up to the total number of flows observed in a system. This leads to this particular model being more accurately described as a 'total constrained' model, where:

$$k = \frac{T}{\sum^i \sum^j V_i^{\mu} W_j^{\alpha} d_{ij}^{-}\beta} \tag{3}$$

and T is the sum of our matrix of observed flows or:

$$T = \sum_i \sum_j T_{ij} \tag{4}$$

In plain language, this is just the sum of all observed flows divided by the sum of all of the other elements in the model.

If we use distance as a basic measure of cost, then the simplest distance to measure is the Euclidean distance between the centroids of the zones for which you have data. In the accompanying walk-through exercise, the process of generating a distance matrix from a set of GCCSA boundaries is demonstrated using the `spdists()` function in R.

Observing equations 1 and 2 above you would be forgiven for asking how the values for the four parameters are known or estimated. One solution could be to simply insert some arbitrary or expected values as parameters. This becomes more feasible once we understand what the parameters are in reality: they relate to the scaling effect/importance of the variables with which they are associated. Most simply, where the effects of origin and destination attributes on flows scale in a linear fashion (i.e. for a 1 unit increase in, say, population at origin, we might expect a 1 unit increase in flows of people from that origin; or for a halving in average salary at destination, we might expect a halving of migrants), then the parameter/scaling factor for our origin variable would equal 1, e.g. μ = 1 and α = 1.

In Newton's original gravity equation the negative influence of distance is not linear. Rather, it follows a power law. For example, where β = -2 for a 1 unit increase in distance, we have a 1^{-2} (or 1) unit decrease in interaction/flow. For a 2 unit increase in distance, we have 2^{-2} (0.25 or 1/4) of the interaction, for a 3 unit increase, 3^{-2} (0.111) of the interaction and so on.

We can check to see if μ = 1 and α = 1 and β = -2 are a good or poor guess by looking at our data and plotting observed flows against each variable and seeing whether the value of these variables raise to these powers (Figure 2). Reviewing the three graphs at Figure 2a, 2b, 2c, it can be observed that -2 looks like a fairly good estimate for β with the red modelled line matching quite closely the observed relationship between migration flows and distances. 1 also looks like a fair estimate of μ for at least some of the relationship between origin population and migration flows; however, it appears that there is little discernible visible relationship between destination median incomes and migration flows, so the value of α may be of little consequence.

If we accept that these first parameter estimates are plausible, then they can be inserted into equation 2 to generate a first set of estimates. Such a set of estimates, where k=3.28, μ = 1, α=1 and β=-2, are shown in Table 2. These can be compared with the observed flows in Table 3.

Manual inspection of the flows reveals that in some cases the estimates are not too far from the observed flows, but in others we can clearly see that the estimates are a long way out. Whilst it is OK to 'eyeball' small flow matrices like these, when you have much larger matrices, another solution is required to test the so-called 'goodness-of-fit'. There are a number of ways to do this but two of the most common are to calculate the coefficient of determination (R^2) or the Square Root of Mean Squared Error (RMSE). Anyone who has run a linear regression model before will have come across R^2 but RMSE may be less familiar. There are other methods and they all do more-or-less the same thing: compare the modelled estimates with the real data and represent the degree of agreement with a single number. R^2 is popular as it is quite intuitive and can be compared across models. RMSE is less intuitive, but some argue is better for comparing changes to the same model.

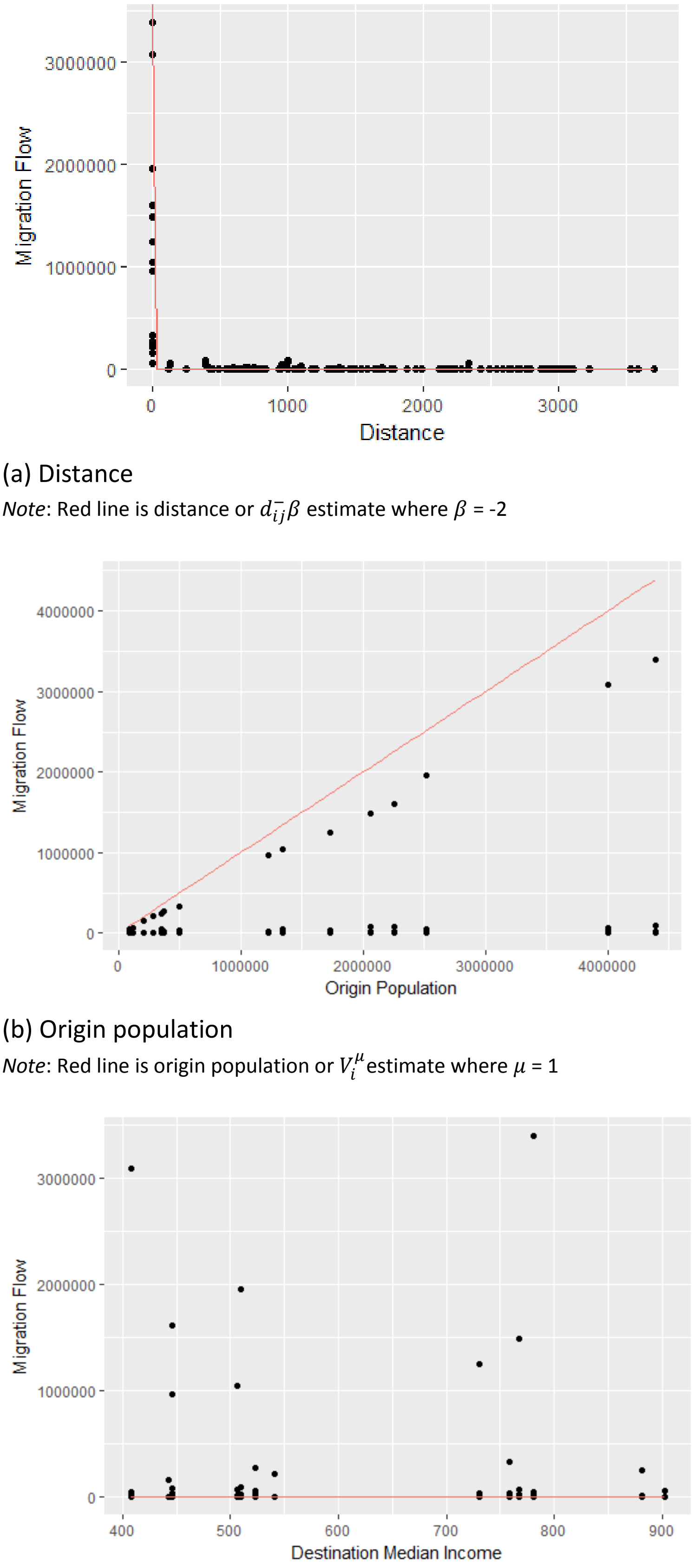
Guidance on how to quickly calculate R^2 and RMSE can be found in the accompanying practical guide. In this initial case, the R^2 value is 0.18. This tells us that this first model accounts for about 18 per cent of the variation of flows in the system – not brilliant, but a starting point nevertheless.

As a result, two immediate questions emerge:

Can we improve these estimates?

Can we tell which predictor variables are best?

Fortunately, the answer to both of these questions is 'yes'. One way that we can begin to answer both of these questions is through the process of model calibration.



(a) Distance

Note: Red line is distance or $d_{ij}^{-}\beta$ estimate where β = -2

(b) Origin population

Note: Red line is origin population or V_i^{μ}estimate where μ = 1

(c) Destination median income

Note: Red line is destination median income or W_j^{α}estimate where α = 1

Figure 2: The relationship between migration flows and three predictor variables

Table 2: Modelled flows from the initial total constrained gravity model with crude estimated parameters

Origin / Destination	1GSYD	1RNSW	2GMEL	2RVIC	3GBRI	3RQLD	4GADE	4RSAU	5GPER	5RWAU	6GHOB	6RTAS	7GDAR	7RNTE	8ACTE	(all)
1GSYD	0	47944	12607	15513	22050	3346	5187	3522	1012	1325	5627	7789	1386	1395	193311	322014
1RNSW	41995	0	8089	12786	11218	3039	5467	3675	721	1008	2566	3703	1021	1138	42659	139085
2GMEL	21972	16095	0	373099	5627	2040	13506	4665	1294	1539	15973	29621	1221	1271	62311	550234
2RVIC	7325	6893	101083	0	2014	784	6705	1971	465	570	3652	6304	448	481	18995	157690
3GBRI	10556	6131	1546	2041	0	3035	1256	1293	397	573	955	1259	791	771	6639	37243
3RQLD	3002	3113	1050	1491	5688	0	1508	2718	615	1118	630	844	2101	2759	2873	29510
4GADE	2536	3051	3788	6941	1283	821	0	5788	653	858	1300	1973	543	657	4041	34233
4RSAU	441	525	335	523	338	379	1483	0	250	462	172	241	275	439	569	6432
5GPER	425	346	312	414	349	288	562	839	0	4637	273	356	730	604	535	10670
5RWAU	156	136	104	142	141	147	207	434	1299	0	81	107	523	508	190	4175
6GHOB	478	249	778	657	169	60	226	116	55	58	0	25438	46	43	876	29249
6RTAS	724	393	1579	1240	244	87	375	179	79	84	27825	0	65	62	1416	34352
7GDAR	33	28	17	23	39	56	26	52	41	105	13	17	0	326	36	812
7RNTE	42	39	22	31	49	93	41	107	44	131	15	20	417	0	47	1098
8ACTE	13901	3502	2571	2893	997	230	594	327	92	116	742	1096	110	112	0	27283
(all)	103586	88445	133881	417794	50206	14405	37143	25686	7017	12584	59824	78768	9677	10566	334498	1384080

Source: ABS 2011 Census. *Notes*: Greater Sydney = 1GSYD; Rest of New South Wales = 1RNSW; Greater Melbourne = 2GMEL; Rest of Victoria = 2RVIC; Greater Brisbane = 3GBRI; Rest of Queensland = 3RQLD; Greater Adelaide = 4GADE; Rest of South Australia = 4RSAU; Greater Perth = 5GPER; Rest of Western Australia = 5RWAU; Greater Hobart = 6GHOB; Rest of Tasmania = 6RTAS; Greater Darwin = 7GDAR; Rest of Northern Territories = 7RNTE; Australian Capital Territory = 8ACTE

Table 3: Original flow data for comparison

Origin / Destination	1GSYD	1RNSW	2GMEL	2RVIC	3GBRI	3RQLD	4GADE	4RSAU	5GPER	5RWAU	6GHOB	6RTAS	7GDAR	7RNTE	8ACTE	(all)
1GSYD	0	91043	22605	4420	22874	27447	5829	795	10572	2127	1654	1984	1992	828	10658	204828
1RNSW	53568	0	12418	13072	21289	35191	3613	1587	4999	3295	978	1885	2252	1431	15766	171344
2GMEL	15569	11094	0	70264	13055	16164	6017	1292	10111	2570	2126	2553	2029	1004	4727	158575
2RVIC	2528	11968	47988	0	4328	10110	3468	2217	3449	2597	667	1428	1548	721	1362	94379
3GBRI	12333	16056	13080	4249	0	84649	3044	818	4810	1796	1388	2295	1802	905	3127	150352
3RQLD	11629	26699	12284	7566	74412	0	3772	1758	6583	4688	1479	3086	3126	2142	3125	162349
4GADE	5415	3517	8803	3188	5449	6178	0	25679	3831	1230	598	872	1843	927	1995	69525
4RSAU	477	1490	1154	2439	824	2631	22020	0	1051	1354	148	429	679	484	185	35365
5GPER	6523	4064	11721	2931	5086	7019	2625	865	0	41332	1022	1802	1305	416	1675	88386
5RWAU	714	2241	1490	1811	1141	4333	808	982	42149	0	277	1161	1093	627	251	59078
6GHOB	1221	998	3014	624	1307	1810	532	111	901	365	0	5019	195	113	564	16774
6RTAS	1029	1871	2637	1647	1543	2884	658	343	1210	1028	7214	0	272	164	288	22788
7GDAR	1237	2185	1957	1481	2763	5107	2111	641	2149	949	239	333	0	1998	824	23974
7RNTE	406	1432	700	792	896	3018	1296	961	699	826	96	213	2684	0	229	14248
8ACTE	7065	16829	5930	1994	5225	6968	2657	1091	2212	1110	466	480	3304	56779	0	112110
(all)	119714	191487	145781	116478	160192	213509	58450	39140	94726	65267	18352	23540	24124	68539	44776	1384075

Source: ABS 2011 Census

2.2.2. Regression modelling framework

Calibration is the process of adjusting parameters in the model to try and get the estimates to agree with the observed data as much as possible. Adjusting the parameters is the sort of iterative process that computers are particularly good at and the goodness-of-fit statistics can be used to indicate when the optimum solution is found. Historically this process required a researcher with the requisite programming skills to write a computer algorithm to iteratively adjust each parameter, check the goodness-of-fit, and then start all over again until the goodness-of-fit statistic was maximised/minimised. There are various well-established routines that can achieve this, such as the Newton-Raphson algorithm, but without the necessary programming skills this can be a serious barrier and probably why spatial interaction modelling was the preserve of a few specialists for so long.

However, since the early days of spatial interaction modelling, a number of useful developments have occurred. Perhaps the most important in the context of calibration is the fact that it is possible to turn the multiplicative model in equation 2 into an additive model. Taking the logarithms of both sides of equation 2, you end up with the following equation:

$$\ln T_{ij} = k + \mu \ln V_i + \alpha \ln W_j - \beta \ln d_{ij} \tag{5}$$

What we have now is a regression model. Anyone who has been introduced to regression models in introductory statistics classes will be aware that there are various pieces of software available to us to run regressions (such as R) and calibrate the parameters (or ‘estimate the coefficients’ in the language of statistics), so expert programming skills are no longer required.

There are some papers that are worth reading at this point if you would like to learn more. Perhaps the best is by Flowerdew and Aitkin (1982). One of the key points that Flowerdew and Aitkin make is that the model in equation 5 (known as a log-normal model) has various problems associated with it, which mean that the estimates produced might not be reliable. The paper (and also Wilson’s 1971 paper) details these issues; however, the salient point is that the way around many of these issues is to re-specify the model, not as a log-normal regression but as a Poisson or negative binomial regression model.

The flows that spatial interaction models deal with (such as migration or commuting) relate to non-negative integer counts (you cannot have negative people moving between places and you cannot normally – if they are alive! – have fractions of people moving either). As such, the probability of migrating or commuting is not described by a continuous (normal) probability distribution (the distribution which underpins the error distribution in standard linear regression models), but a discrete probability distribution such as the Poisson distribution or the negative binomial distribution (of which the Poisson distribution is a special case).

There is a family of generalised linear models, but in the analysis of migration and other population flows Poisson and Negative Binomial Regression models have been used most frequently (Abel 2010; Congdon 1993, 1988; Crymble, Dennett and Hitchcock 2017; Flowerdew 2010, 1982; Flowerdew and Aitkin 1982; Shen 2017, 2015; Willekens 1999). The differences between the two are technical, but some, including Congdon (1993), argue that Negative Binomial models should be used over Poisson due to a statistical phenomenon known as overdispersal. For ease of explanation here we will continue with the Poisson model, but be aware that in practice a Negative Binomial model may be better.

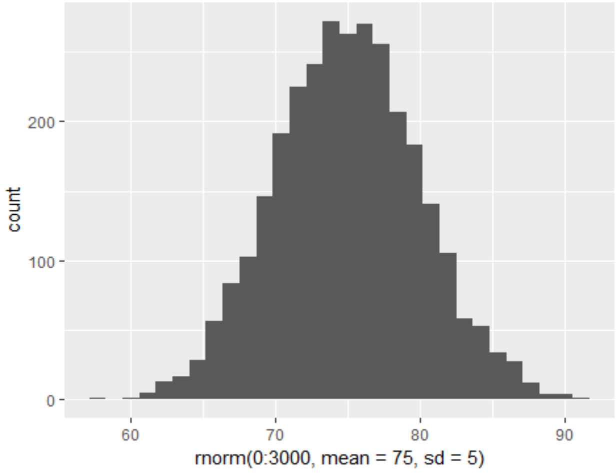
In the migration modelling literature others, such as Raymer (Raymer 2007; Raymer, Abel and Smith 2007; Raymer and Giulietti 2010; Rogers and Raymer 1998), have used what are described as log-linear models. These are exactly the same as Poisson models with the distinction that Poisson models will usually contain both continuous and categorical predictor variables, whereas log-linear models will only contain categorical predictors. For a two-dimensional population flow matrix between origins and destinations, these categorical predictors would generally be the origin and destination zones – in effect we would have a two-dimensional contingency table. The analysis of contingency tables is well established in statistics and as such has its own lexicon. In log-linear modelling terminology, these origin and destination zones would be described as the ‘main’ or ‘fixed’ effects and are equivalent to the ‘constraints’ that will be introduced later on in this paper.

This discussion of Poisson, log-linear and spatial interaction models is included here for the purpose of highlighting that in the migration modelling literature it can be particularly confusing when reading papers by different authors who all use very different terminology and modelling paradigms. Iterative Proportional Fitting (IPF) is another term that may appear when researching papers in this area (Lomax and Norman 2016). The salient point is that all of the models are essentially doing the same thing, but the papers in which they are outlined ascribe to different definitional conventions.

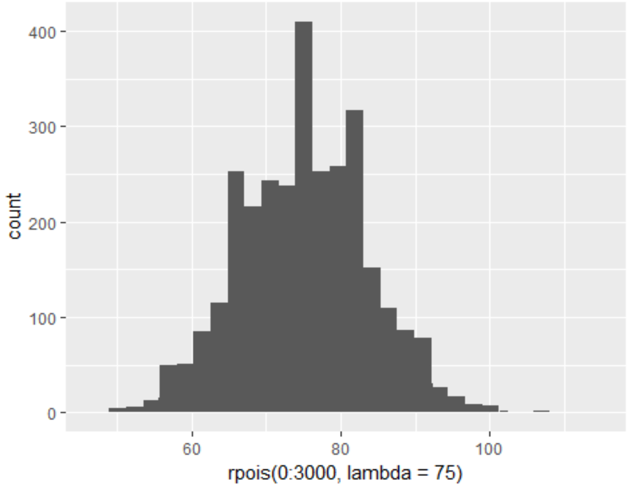
How is the Poisson distribution different to a normal distribution? Aside from them describing different frequency/probability distributions, they behave differently for different sets of observations. Below are two histograms (Figure 3). The first is a random variable with a normal distribution, with a mean of 75 and a standard deviation of 5; the second is a histogram of Poisson distributed variable with the same mean (Poisson distributions have only one parameter – the mean).

You will notice that they look broadly similar. However, with a Poisson distributed variable, when the mean (λ - lambda) changes, so does the shape of the frequency distribution. As the mean gets smaller, and this is often the case with flow data where small flows are very likely, the distribution starts to look a lot more like a skewed or log-normal continuous distribution. The key point is that it is not a continuous distribution but a discrete (Poisson) distribution. Figure 4 plots a frequency distribution for a discrete variable with a small mean. The shape of the histogram is very similar to a positively skewed continuous distribution.

For any system of population flows between a matrix of origins and destinations, the flows will have a mean value of λ_{ij} which will normally be quite low and will dictate the distribution. Plotting the frequency distribution of migration flows between our Australian GCCSAs (excluding within-area flows – Figure 5), reveals a histogram which looks like a skewed normal or, more accurately, a Poisson distribution with a small mean.



(a) Random variable with a normal distribution



(b) Random variable with a Poisson distribution

Figure 3: Normal and Poisson distributions

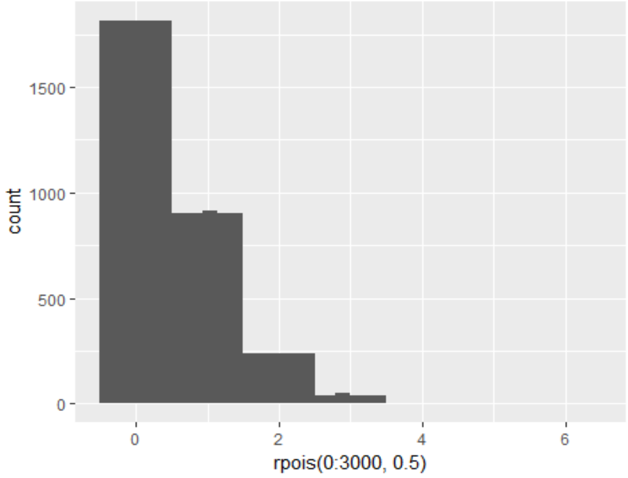


Figure 4: Poisson distribution with a small mean

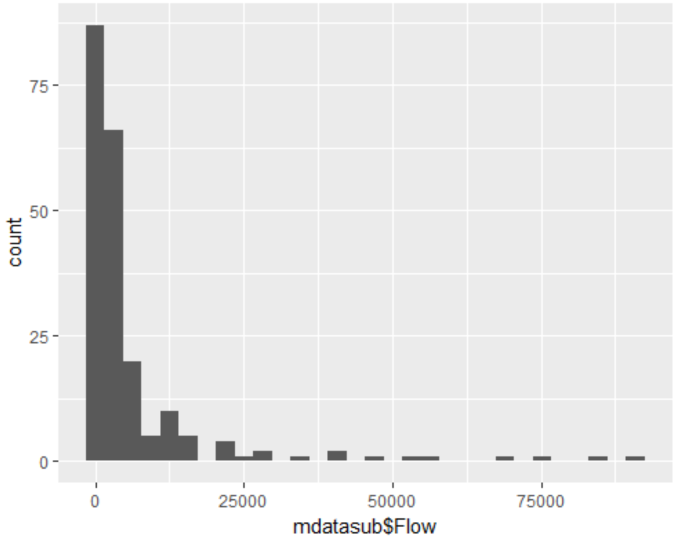


Figure 5: Frequency distribution of migration flows (excluding intra-zonal flows) between GCCSAs

In all of this discussion about frequency/probability distributions, it can be easy to lose track of the purpose for understanding all of this. Perhaps the easiest way to remind ourselves of the purpose is in understanding the basics of what a regression model is trying to do. At its most simple a regression model is nothing more than a line of best fit drawn through a cloud of observations. If we think of a spatial interaction model as partially representing the relationship between volume of flow and cost of interaction (distance), then we would expect to see a straight line between flow volumes and distance. Figure 2a shows that when plotting raw flows against distance, the relationship cannot be represented by a straight line. However, if both the flows and the distance are logged, as in equation 5, a plot similar to the one below in Figure 6 is produced. It might not be a clear straight-line relationship, but there is certainly a suggestion that the blue regression line does represent the underlying relationship between migration flows and distance.

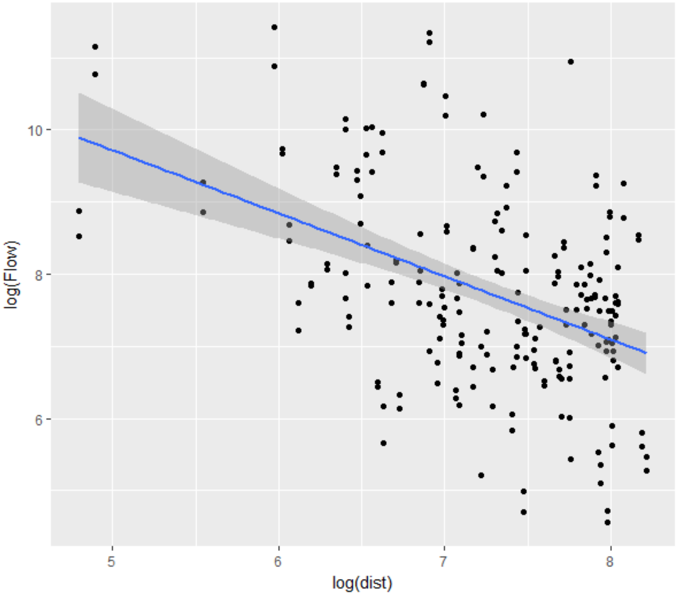


Figure 6: Relationship between log(distance) and log(migration flows) between GCCSAs

Now the discussion in the previous session indicates that the y variable in our model is not logged as in the graph above. However, it can still be modelled using something like the blue line if we assume a Poisson distribution.

Equation 5 can now be re-specified as a Poisson regression model. Instead of the dependent variable being $\ln T_{ij}$, it is now the mean of the Poisson distribution λ_{ij} and the model becomes:

$$\lambda_{ij} = \exp(k + \mu \ln V_i + \alpha \ln W_j - \beta \ln d_{ij}) \qquad (6)$$

What this model says is λ_{ij} (the dependent variable – the estimate of T_{ij}) is *logarithmically linked* to (or modelled by) a linear combination of the logged independent variables in the model. Using equation 6, a Poisson regression model can be fitted to produce estimates of k, μ, α and β – or put another way, we can use the regression model to calibrate our parameters.

It is very straight forward to run a Poisson regression model in R using the `glm()` (Generalised Linear Models) function. The code to do this can be found in the accompanying guide. Delving into the depths of the `glm()` function documentation will reveal that the parameters are calibrated though an ‘iteratively re-weighted least squares’ algorithm. This essentially fits lots of lines similar to that in Figure 6 to the data until it finds the best one. It continually adjusts the parameters to minimise the error between the observed and expected (blue line) values using some goodness-of-fit measure, not dissimilar to an R^2 or RMSE. Running the model will produce some output similar to that shown below in Box 1.

```
Call:
glm(formula = Flow ~ log(vi1_origpop) + log(wj3_destmedinc) +
    log(dist), family = poisson(link = "log"), data = mdatasub,
    na.action = na.exclude)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-177.78   -54.49   -24.50     9.21   470.11

Coefficients:
                      Estimate Std. Error z value             Pr(>|z|)
(Intercept)          7.1953790  0.0248852  289.14 <0.0000000000000002 ***
log(vi1_origpop)     0.5903363  0.0009232  639.42 <0.0000000000000002 ***
log(wj3_destmedinc) -0.1671417  0.0033663  -49.65 <0.0000000000000002 ***
log(dist)           -0.8119316  0.0010157 -799.41 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2750417  on 209  degrees of freedom
Residual deviance: 1503573  on 206  degrees of freedom
AIC: 1505580

Number of Fisher Scoring iterations: 5
```

Box 1: Model output including calibrate parameters from R glm() implementation of equation 6

Box 1 contains various pieces of information. The 'Call' section at the top is the code used to run the model in R. Then, under the 'Coefficients' section, are the values for the four calibrated parameters in the model. In our original model, we estimated the four parameters as follows:

$k = 3.28$, $\mu = 1$, $\alpha = 1$ and $\beta = -2$

After fitting the Poisson model, the values for the parameters can be found in the 'Estimate' column. They change to:

$k = 7.195$, $\mu = 0.59$, $\alpha = -0.17$ and $\beta = -0.81$

The regression model also produces some other useful pieces of output. The p-values in the last column reveal that all variables have a highly (***) statistically significant influence on migration flows, with the z-scores (standardised coefficients) revealing that distance has the most (negative) influence on the model followed by origin population, with destination income only a small influence on the flows, and a counterintuitive one at that, with increases in destination income resulting in decreases in migration flows.

In the accompanying practical guide, it is shown how these parameter values can be inserted directly back into equation 6 to produce a new set of estimates similar to those in Table 1. The R^2 value for this new matrix improves to 0.32, meaning that by simply calibrating the model parameters on observed data, we are able to explain around 14 per cent more of the variation in the migration flows in our system.

2.2. Constrained spatial interaction models

Returning to Wilson's (1971) seminal paper, he introduces a full *family* of spatial interaction models of which the unconstrained model is just the start. Of course, since then, there have been all manner of incremental advances and alternatives (Dennett and Wilson 2013; Fotheringham 1983; Pooler 1994; Stillwell 1978). However, in this section we will concentrate on Wilson's original family – the Production (origin) Constrained Model; the Attraction (destination) Constrained Model; and the Doubly Constrained Model – but show how the Poisson regression framework can be used as with the unconstrained model.

Recalling the unconstrained/total constrained model above (Table 1), while the total flows in the estimates equalled the total observed flows, none of the estimates sum to the observed in-migration and out-migration totals (the margins of the matrix). Wilson's real contribution to the field was in noticing that this unconstrained model was sub-optimal as it did not make use of all of the available information in the system being studied.

Where there is a full flow matrix to calibrate parameters, then it is possible to incorporate the row (origin) totals, column (destination) totals or both origin and destination totals to constrain flow estimates to these known values. There are various reasons for wanting to do this in different flow modelling contexts, for example:

a) If the researcher is interested in flows of money into businesses or customers into shops, then they might have information on the amount of disposable income and shopping habits of the people living in different areas, perhaps from loyalty card data. This is known information about origins and so it would be logical to constrain the estimates from a spatial interaction model to this known information. Other information about the attractiveness of shops and businesses (store size, variety/specialism of goods etc.) can then be used to estimate how much money/customers a new store opening in the area might make/attract or, if a new out-of-town shopping centre opens, how much it might affect the business of shops in the town centre. This is what is known in the literature as the 'retail model' and is perhaps the most common example of a Production (origin) Constrained Spatial Interaction Model.

b) Other researchers might be interested in understanding the impact of a large new employer in an area on the flows of traffic in the vicinity or on the demand for new worker accommodation nearby. A good example of where this might be the case is with large new infrastructure developments like airports. For example, before the go-ahead for the new third runway at Heathrow Airport in London, England was given, one option being considered was a new runway in the Thames Estuary. If a new airport was built here, what would the potential impact on transport flows be in the area and where might workers commute from? This sort of scenario could be tested with an Attraction (destination) Constrained Spatial Interaction Model where the number of new jobs in a destination is known (as well as jobs in the surrounding area). The model could also be used to estimate where the workers will be drawn from and their likely travel-to-work patterns. These models are known as Land Use Transport Interaction (LUTI) models and have a well-established history in urban planning.

c) Other researchers might be interested in understanding the changing patterns of commuting or migration over time. Data from a census provides an accurate snap-shot of migrating and commuting patterns, but only periodically. In these full data matrices, information about both the numbers of commuters/migrants leaving origins and arriving at destinations and the interactions between them is known. Constraining model estimates to this known information at origin and destination allows various things to be examined, including:

   i. the ways that the patterns of commuting/migration differ from the model predictions – where might there be more migrant/commuter flows than expected?

   ii. how the model parameters vary over time – for example, how does distance/cost of travel affect flows over time? Are people prepared to travel further or less distance than before?

2.2.2. The Production Constrained Model

Recall the unconstrained model from equation 2. A Production Constrained Model constrains estimates to known information about the origins and so replaces the terms k and V_i^{μ} to produce the following model:

$$T_{ij} = A_i O_i W_j^{\alpha} d_{ij}^{-} \beta \qquad (7)$$

where:

$$O_i = \sum_j T_{ij} \qquad (8)$$

and:

$$A_i = \frac{1}{\sum_j W_j^{\alpha} d_{ij}^{-} \beta} \qquad (9)$$

In the Production Constrained Model, O_i does not have a parameter as it is a known constraint. A_i is known as a *balancing factor* and is a vector of values which relate to each origin i which do the equivalent job as k in the unconstrained/total constrained model but ensure that flow estimates from each origin sum to the know totals O_i rather than just the overall total.

Now at this point the O_i and A_i values could be calculated by hand for the sample system and the parameter values for the rest of the model could be guessed. However, the Poisson regression framework allows this to be avoided.

The Production Constrained Model can be re-specified as a Poisson regression model in exactly the same way as before. Taking the logs of the right-hand side of the equation, and assuming that these are logarithmically linked to the Poisson distributed mean (λ_{ij}) of the T_{ij} variable, means that equation 7 becomes:

$$\lambda_{ij} = exp(\mu_i + \alpha \ln W_j - \beta \ln d_{ij}) \tag{10}$$

In equation 10 μ_i is the equivalent of the vector of balancing factors A_i. but in regression/log-linear modelling terminology these can also be described as either 'dummy variables' or 'fixed effects'. In practical terms what this means is that in the regression model μ_i is modelled as a categorical predictor[3], and therefore in the Poisson regression model the numeric values of O_i are ignored and replaced by a categorical identifier for the origin. In terms of the origin/destination migration matrix shown in Table 3, rather than the flow of 204,828 migrants leaving Sydney (row 1) being used as a predictor, simply the code '1GSYD' is used as a dummy variable.

In the accompanying practical guide the code for running this model using the `glm()` function in R is provided. Running the model will produce the following output (Box 2). There are elements of the model output that should be familiar from the unconstrained model:

- The α parameter related to the destination attractiveness (in this case, median weekly income): -0.27 is not much different from the unconstrained model. The z-score indicates that this is not a very important variable in explaining variation in migration behaviours in Australia.
- The β distance decay parameter: -1.23 has decreased meaning after controlling for origin characteristics, distance becomes more of a deterrent.

Where the model output differs is that the intercept (k) parameter has been replaced by the vector of dummy variables/constraints μ_i relating to each origin. We can see from the standard outputs from the model that all of the explanatory variables are statistically significant (***); the z-scores indicate that the rest of Queensland and Greater Sydney have greater emissivity properties than all other zones in the model, with factors associated with these zones more important in explaining migration patterns in Australia than distance, which while still important, is less important than a number of origin zones.

In the accompanying practical guide there is code that allows you to create a new set of estimates by both plugging these values back into a multiplicative model or by using a much easier built-in function in `glm()`. Using either method will produce the set of flows shown in Table 4.

[3] https://en.wikipedia.org/wiki/Categorical_variable

```
Call:
glm(formula = Flow ~ Orig_code + log(wj3_destmedinc) + log(dist) -
    1, family = poisson(link = "log"), data = mdatasub, na.action =
na.exclude)

Deviance Residuals:
    Min        1Q    Median       3Q       Max
-225.71    -54.10    -15.94    20.45    374.27

Coefficients:
                     Estimate Std. Error z value            Pr(>|z|)
Orig_code1GSYD      19.541851   0.023767  822.22 <0.0000000000000002 ***
Orig_code1RNSW      19.425497   0.023913  812.35 <0.0000000000000002 ***
Orig_code2GMEL      18.875763   0.023243  812.12 <0.0000000000000002 ***
Orig_code2RVIC      18.335242   0.022996  797.31 <0.0000000000000002 ***
Orig_code3GBRI      19.856564   0.024063  825.20 <0.0000000000000002 ***
Orig_code3RQLD      20.094898   0.024300  826.94 <0.0000000000000002 ***
Orig_code4GADE      18.747938   0.023966  782.28 <0.0000000000000002 ***
Orig_code4RSAU      18.324029   0.024407  750.75 <0.0000000000000002 ***
Orig_code5GPER      20.010551   0.024631  812.43 <0.0000000000000002 ***
Orig_code5RWAU      19.392751   0.024611  787.96 <0.0000000000000002 ***
Orig_code6GHOB      16.802016   0.024282  691.97 <0.0000000000000002 ***
Orig_code6RTAS      17.013981   0.023587  721.33 <0.0000000000000002 ***
Orig_code7GDAR      18.607483   0.025012  743.93 <0.0000000000000002 ***
Orig_code7RNTE      17.798856   0.025704  692.45 <0.0000000000000002 ***
Orig_code8ACTE      17.796693   0.023895  744.79 <0.0000000000000002 ***
log(wj3_destmedinc) -0.272640   0.003383  -80.59 <0.0000000000000002 ***
log(dist)           -1.227679   0.001400 -876.71 <0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 23087017  on 210  degrees of freedom
Residual deviance:  1207394  on 193  degrees of freedom
AIC: 1209427

Number of Fisher Scoring iterations: 6
```

Box 2: Outputs from the production constrained model specified in equation 10

Comparing Table 4 with Table 3, it is very easy to see the origin constraints working. The sum across all destinations for each origin in the estimated matrix (Table 4) is exactly the same (give or take the odd rounding error) as the same sum across the observed matrix (Table 3): $\sum_j T_{ij} = \sum_j \lambda_{ij} = O_i$. But clearly the same is not true when you sum across all origins for each destination: $\sum_i T_{ij} \neq \sum_i \lambda_{ij} \neq D_j$. Calculating the R^2 value, the fit of the model has improved quite considerably: from around 0.32 in the unconstrained model to around 0.43 in this model. The RMSE has also dropped quite noticeably.

One of the advantages of singly constrained models is that once initial parameters have been calibrated on existing data, then changes can be made to destination variables (for origin/production constrained models) or origin variables (for destination/attraction constrained models) and the impact on flow estimates explored. For example, what would happen if average wages suddenly increased in an area? How would this impact migration flows to and from that area? The accompanying worked example demonstrates how this can be explored.

Table 4: Modelled flows from a Production Constrained Spatial Interaction Model with parameters calibrated by a Poisson regression model

Origin / Destination	1GSYD	1RNSW	2GMEL	2RVIC	3GBRI	3RQLD	4GADE	4RSAU	5GPER	5RWAU	6GHOB	6RTAS	7GDAR	7RNTE	8ACTE	(all)
1GSYD	0	36794	19752	18516	15905	8076	10591	7248	2504	2860	11192	11454	2519	4105	53308	204824
1RNSW	29163	0	18862	20620	13173	9548	13715	9329	2549	3032	8667	9100	2619	4543	26439	171359
2GMEL	8501	10243	0	70950	3742	3243	10367	4685	1584	1705	11552	14147	1268	2109	14474	158570
2RVIC	4924	6918	43838	0	2263	2050	7667	3139	961	1053	5309	6221	779	1320	7935	94377
3GBRI	21684	22658	11852	11604	0	16555	9653	8526	3069	3722	8200	8144	3886	6207	14647	150407
3RQLD	12057	17984	11248	11511	18128	0	12989	16188	4832	6746	7639	7664	8515	16335	10539	162375
4GADE	4109	6714	9345	11186	2747	3376	0	9731	1895	2167	4506	4879	1403	2558	4912	69528
4RSAU	1922	3122	2887	3130	1659	2876	6653	0	1438	2028	1780	1840	1264	2736	2017	35352
5GPER	3930	5048	5777	5673	3533	5080	7666	8507	0	17470	4952	4882	4812	6954	4064	88348
5RWAU	2445	3269	3387	3386	2333	3862	4775	6535	9514	0	2696	2679	4515	7196	2476	59068
6GHOB	619	605	1485	1105	333	283	643	371	175	175	0	9840	129	201	807	16771
6RTAS	827	829	2374	1689	431	371	908	501	225	226	12842	0	166	261	1121	22771
7GDAR	1030	1350	1204	1198	1165	2331	1478	1948	1253	2159	950	937	0	6000	981	23984
7RNTE	644	899	769	779	714	1716	1034	1618	695	1321	569	568	2303	0	618	14247
8ACTE	9622	6021	6070	5386	1939	1274	2285	1373	467	523	2631	2802	433	712	0	41538
(all)	101477	122454	138850	166733	68065	60641	90424	79699	31161	45187	83485	85157	34611	61237	144338	1313519

2.2.3. The Attraction Constrained Model

The Attraction Constrained Model is virtually the same as the Production Constrained Model:

$$T_{ij} = D_j B_j V_i^{\mu} d_{ij}^{-}\beta \qquad (11)$$

where:

$$D_j = \textstyle\sum_i T_{ij} \qquad (12)$$

and:

$$B_j = \frac{1}{\sum_i V_i^{\mu} d_{ij}^{-}\beta} \qquad (13)$$

The Poisson model equation for the Attraction Constrained Model would be:

$$\lambda_{ij} = exp(\mu \ln V_i + \alpha_i - \beta \ln d_{ij}) \qquad (14)$$

Its implementation in R is virtually identical to the Production Constrained Model. See the accompanying walk-through exercise for full details of how to run these models with the sample dataset. Because of the similarities to the Production Constrained Model, the Attraction Constrained Model will not be dwelt upon here.

2.2.3. The Doubly Constrained Model

The final model in the Wilson (1971) family is the Doubly Constrained Model. Let's begin with the formula:

$$T_{ij} = A_i O_i B_j D_j d_{ij}^{-}\beta \qquad (15)$$

where:

$$O_i = \textstyle\sum_j T_{ij} \qquad (16)$$

$$D_j = \textstyle\sum_i T_{ij} \qquad (17)$$

and:

$$A_i = \frac{1}{\sum_j B_j D_j d_{ij}^{-}\beta} \qquad (19)$$

$$B_j = \frac{1}{\sum_i A_i O_i d_{ij}^{-}\beta} \qquad (20)$$

Astute readers will have noticed that the calculation of A_i relies on knowing B_j and the calculation of B_j relies on knowing A_i – something of a conundrum to which the solution is elegantly described by Senior (1979), who sketches out a very useful algorithm for iteratively arriving at values for A_i and B_j by setting each to equal 1 initially and then continuing to calculate each in turn until the difference between successive iterations of the A_i and B_j values is small enough not to matter. In the accompanying practical guide an algorithm to achieve this using this multiplicative framework is provided. However, as you will have probably guessed by now, the Poisson regression framework allows for the Doubly Constrained Model to be fitted very easily.

The Poisson Doubly Constrained Model takes the form:

$$\lambda_{ij} = exp(\mu_i + \alpha_i - \beta \ln d_{ij}) \qquad (21)$$

When run in R and applied to the Australian migration data we have been using, this model will produce the outputs shown in Box 3. The coefficients in this version of the Doubly Constrained Model will look a little different. This is explained in the accompanying walk-through exercise and is because an intercept has been added along with reference categories for the categorical variables, as two factor levels are used in this model. Importantly the model estimates are not altered in any way by this. The reference level means that the origin and destination coefficients need to be interpreted in relation to a reference category. In this example, the first zone in the system is used (Sydney), with the direction and size of the coefficients referring to whether another origin or destination zone has a greater or lesser positive or negative effect on migration flows in the system when compared to Sydney. The estimates produced by the Doubly Constrained Model are the most accurate in the Wilson family of models (in this example, an R^2 value of 0.87). However, there is a loss of flexibility when compared to the singly constrained models, as only alternatives to origin/destination interaction explanatory variables such as historic flows or something other than distance can be experimented with. The double constraints mean that origin- and destination-specific explanatory variables cannot be used.

2.2.4. Further experimentation

All of the way through this paper there has been an assumption that the distance decay parameter follows a negative power law. This does not have to be the case and empirically might not necessarily be so. In his original paper, Wilson (1971) generalised the distance decay parameter to:

$$f(d_{ij}) \qquad (22)$$

where f represents some function of distance describing the rate at which the flow interactions change as distance increases. Lots of people have written about this, including Taylor (1983) and more recently Lovelace (2015) in a transport context. The inverse power law that has been used to this point is one possible function of distance; the other common one that is used is the negative exponential function:

$$exp(-\beta d_{ij}) \qquad (23)$$

The exact effect that the different function has on the rate of distance decay will depend on the value of β as well as the function. However, Figure 7 shows how the different functions and values of β can combine to affect distance decay.

```
Call:
glm(formula = Flow ~ Orig_code + Dest_code + log(dist), family = 
poisson(link = "log"), 
    data = mdatasub, na.action = na.exclude)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-93.018  -26.703    0.021   19.046  184.179  

Coefficients:
               Estimate Std. Error  z value            Pr(>|z|)    
(Intercept)    20.208178   0.011308 1786.999 <0.0000000000000002 ***
Orig_code1RNSW -0.122417   0.003463  -35.353 <0.0000000000000002 ***
Orig_code2GMEL -0.455872   0.003741 -121.852 <0.0000000000000002 ***
Orig_code2RVIC -1.434386   0.004511 -317.969 <0.0000000000000002 ***
Orig_code3GBRI  0.241303   0.003597   67.091 <0.0000000000000002 ***
Orig_code3RQLD  0.772753   0.003599  214.700 <0.0000000000000002 ***
Orig_code4GADE -0.674261   0.004527 -148.936 <0.0000000000000002 ***
Orig_code4RSAU -1.248974   0.005889 -212.091 <0.0000000000000002 ***
Orig_code5GPER  0.742687   0.004668  159.118 <0.0000000000000002 ***
Orig_code5RWAU -0.317806   0.005131  -61.943 <0.0000000000000002 ***
Orig_code6GHOB -2.270736   0.008576 -264.767 <0.0000000000000002 ***
Orig_code6RTAS -1.988784   0.007477 -265.981 <0.0000000000000002 ***
Orig_code7GDAR -0.797620   0.007089 -112.513 <0.0000000000000002 ***
Orig_code7RNTE -1.893522   0.008806 -215.022 <0.0000000000000002 ***
Orig_code8ACTE -1.921309   0.005511 -348.631 <0.0000000000000002 ***
Dest_code1RNSW  0.389478   0.003899   99.894 <0.0000000000000002 ***
Dest_code2GMEL -0.007616   0.004244   -1.794              0.0727 .  
Dest_code2RVIC -0.781258   0.004654 -167.854 <0.0000000000000002 ***
Dest_code3GBRI  0.795909   0.004037  197.178 <0.0000000000000002 ***
Dest_code3RQLD  1.516186   0.003918  386.955 <0.0000000000000002 ***
Dest_code4GADE -0.331189   0.005232  -63.304 <0.0000000000000002 ***
Dest_code4RSAU -0.627202   0.006032 -103.980 <0.0000000000000002 ***
Dest_code5GPER  1.390114   0.005022  276.811 <0.0000000000000002 ***
Dest_code5RWAU  0.367314   0.005362   68.509 <0.0000000000000002 ***
Dest_code6GHOB -1.685934   0.008478 -198.859 <0.0000000000000002 ***
Dest_code6RTAS -1.454819   0.007612 -191.112 <0.0000000000000002 ***
Dest_code7GDAR -0.308516   0.007716  -39.986 <0.0000000000000002 ***
Dest_code7RNTE -1.462020   0.009743 -150.060 <0.0000000000000002 ***
Dest_code8ACTE -1.506283   0.005709 -263.866 <0.0000000000000002 ***
log(dist)      -1.589102   0.001685 -942.842 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2750417  on 209  degrees of freedom
Residual deviance:  335759  on 180  degrees of freedom
AIC: 337818

Number of Fisher Scoring iterations: 6
```

Box 3: Outputs from the doubly constrained model specified in equation 21

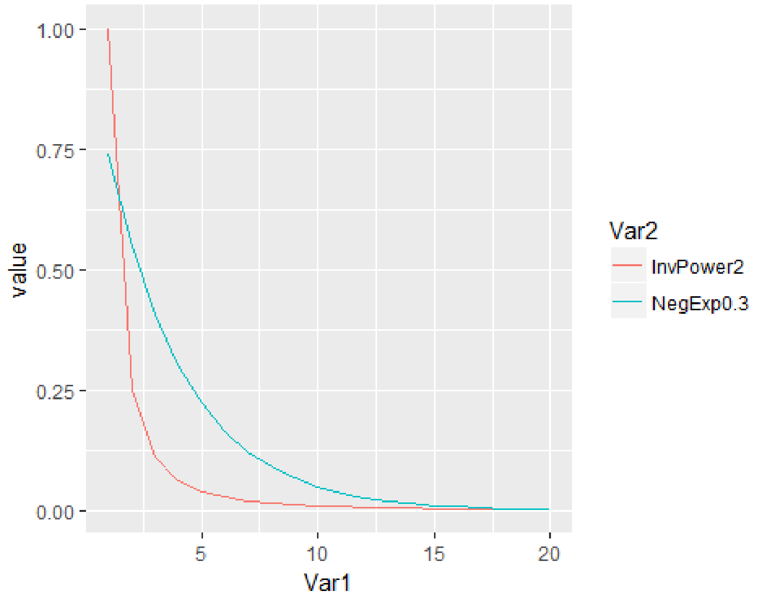


Figure 7: Alternative distance decay curves for alternative values of β and different functions

In this particular example with these parameters and β values, the inverse power function has a far more rapid distance decay effect than the negative exponential function. In real life, what this means is that if the observed interactions drop off very rapidly with distance, then they might be more likely to follow an inverse power law. This might be the case when looking at trips to the local convenience store by walking, for example. On the other hand, if the effect of distance is less severe – for example, migration across the country for a new job – then the negative exponential function with a small value of β function might be more appropriate. There is no hard and fast rule as to which function to pick. It will just come down to which fits the data better. Following Oshan's (2016) example, the accompanying walk-through exercise will allow you to explore the effect on model fits and predictions of fitting different distance decay functions to the distance variable.

The final point to note is that the regression modelling framework means that adding additional explanatory variables into the spatial interaction model is very easy compared with the multiplicative framework. In the accompanying walk-through guide, in addition to variables relating to median income, there are variables on unemployment rate and the percentage of households living in rented accommodation. Experiment with these variables for origins and destinations to see whether the singly constrained models can be improved in any way.

3. Conclusions

Spatial Interaction Modelling is one of the key tools in the population geographer's tool kit, but for too long has been inaccessible to researchers new to the field or without computer programming expertise. Recent advances in more forgiving software environments like R and Python now mean that with much less (although admittedly still some) effort, this powerful modelling tool can be accessed by more people. This guide has been designed to introduce researchers to spatial interaction modelling in, hopefully, an accessible way through exemplification of two modelling frameworks – the Wilson-esque multiplicative framework and the Poisson Regression additive modelling framework – and an accompanying walk-through guide.

Acknowledgements

Thanks to Hadrien Salat in CASA who helped with some earlier technical errors in the modelling.

References

Abel G J (2010) Estimation of international migration flow tables in Europe: international migration flow tables. *Journal of the Royal Statistical Society*. Series A, Statistics in Society 173(4): 797–825.

Congdon P (1993) Approaches to modelling overdispersion in the analysis of migration. *Environment and Planning A: Economy and Space* 25(1): 1481–1510.

Congdon P (1988) Modelling migration flows between areas: an analysis for London using the census and OPCS Longitudinal Study. *Regional Studies* 23(2): 87–103.

Crymble A, Dennett A and Hitchcock T (2017) Modelling regional imbalances in English plebeian migration to late eighteenth-century London. *The Economic History Review* 71(3): 747-771.

Dennett A and Wilson A (2013) A multi-level spatial interaction modelling framework for estimating inter-regional migration in Europe. *Environment and Planning A: Economy and Space* 45(6): 1491–1507.

Erhardt G D and Dennett A (2017) Understanding the role and relevance of the census in a changing transportation data landscape. Paper presented at the Transportation Research Board Conference on Applying Census Data for Transportation, Kansas City, Missouri.

Flowerdew R (2010) Modelling migration with Poisson regression. In: Stillwell J, Duke-Williams O and Dennett A (eds) *Technologies for Migration and Commuting Analysis: Spatial Interaction Data Applications*. Hershey PA: IGI Global.

Flowerdew R (1982) Fitting the lognormal gravity model to heteroscedastic data. *Geographical Analysis* 14(3): 263–267.

Flowerdew R and Aitkin M (1982) A method of fitting the gravity model based on the Poisson distribution. *Journal of Regional Science* 22(2): 191–202.

Fotheringham A S (1983) A new set of spatial-interaction models: the theory of competing destinations. *Environment and Planning A: Economy and Space* 15(1): 15–36.

Fotheringham A S, Nakaya T, Yano K, Openshaw S and Ishikawa Y (2001) Hierarchical destination choice and spatial interaction modelling: a simulation experiment. *Environment and Planning A: Economy and Space* 33(5): 901–920.

Kim K and Cohen J E (2010) Determinants of international migration flows to and from industrialized countries: a panel data approach beyond gravity. *International Migration Review* 44(4): 899–932.

Lee E S (1966) A theory of migration. *Demography* 3(1): 47–57.

Lomax N and Norman P (2016) Estimating population attribute values in a table: “get me started in” iterative proportional fitting. *The Professional Geographer* 68(3): 451–461.

Lovelace R (2015) Estimating distance decay for the national propensity to cycle tool. https://www.slideshare.net/ITSLeeds/estimating-distance-decay-for-the-national-propensity-to-cycle-tool.

Oshan T M (2016) A primer for working with the Spatial Interaction modeling (SpInt) module in the python spatial analysis library (PySAL). *REGION* 3: R11–R23.

Pooler J (1994) An extended family of spatial interaction models. *Progress in Human Geography* 18(1): 17–39.

Pooler J (1987) Modeling interprovincial migration using entropy-maximizing methods. *The Canadian Geographer* 31(1): 57–64.

Raymer J (2007) The estimation of international migration flows: a general technique focused on the origin–destination association structure. *Environment and Planning A: Economy and Space* 39(4): 985–995.

Raymer J and Abel G (2008) Methods to improve estimates of migration flows – the MIMOSA model for estimating international migration flows in the European Union. UNECE/Eurostat work session on migration statistics, Working Paper 8, Geneva.

Raymer J, Abel G and Smith P W (2007) Combining census and registration data to estimate detailed elderly migration flows in England and Wales. *Journal of the Royal Statistical Society*. Series A, Statistics in Society 170(4): 891–908.

Raymer J and Giulietti C (2010) Analysing structures of interregional migration in England. In: Stillwell J, Duke-Williams O and Dennett A (eds) *Technologies for Migration and Commuting Analysis: Spatial Interaction Data Applications*. Hershey PA: IGI Global.

Rees P (1977) The measurement of migration from census data and other sources. *Environment and Planning A: Economy and Space* 9(3): 257–280.

Rogers A and Raymer J (1998) The spatial focus of US interstate migration flows. *Population, Space and Place* 4(1): 63–80.

Senior M L (1979) From gravity modelling to entropy maximizing: a pedagogic guide. *Progress in Human Geography* 3(2): 175–210.

Shen J (2017) Modelling interregional migration in China in 2005–2010: the roles of regional attributes and spatial interaction effects in modelling error. *Population, Space and Place* 23(3): e2014.

Shen J (2015) Explaining interregional migration changes in China, 1985–2000, using a decomposition approach. *Regional Studies* 49(7): 1176–1192.

Stillwell J (1978) Interzonal migration: some historical tests of spatial-interaction models. *Environment and Planning A: Economy and Space* 10(1): 1187–1200.

Taylor P J (1983) *Distance decay in spatial interactions*. Norwich: Geo Books.

Willekens F (1999) Modeling approaches to the indirect estimation of migration flows: from entropy to EM. *Mathematical Population Studies* 7(3): 239–278.

Wilson A (1971) A family of spatial interaction models, and associated developments. *Environment and Planning A: Economy and Space* 3(1): 1–32.

Zipf G K (1946) The P1 P2 / D hypothesis: on the intercity movement of persons. *American Sociological Review* 11(6): 677–686.