Quick and Easy Recipes for Healthy Migration Research

Boyana Buyuklieva*[1] and Adam Dennett†[1]

[1]Bartlett Centre for Advanced Spatial Analysis, UCL

January 12, 2017

Summary

Migration is a time-relevant topic and an interesting example of an interdisciplinary field that has been dabbled in by GI Scientists but where stakeholders, by definition, come from different backgrounds and conventions. The growth in the popularity of Open Source software and open code repositories mean that methods once obscured from branches of the field through technical jargon can now be easily accessed. This paper will re-examine several 'quick and easy' ways to quantify migration. These are metrics that require the least amount of raw data to compute and can therefore be considered as a useful starting point for migration research. Firstly, the paper will provide a review of the discussion on measuring migration and the literate programming paradigm. The follow sections will: focus on creating a visual explanation of the input data-set, express formulas using generic language-independent vector arithmetic and show an example of how ambiguities can yield different results.

KEYWORDS: geo-computation, reproducibility, literate programming, interdisciplinary, quantitative migration

1. Introduction

Advancing any field requires a common foundation of reproducible research. The increasing growth of new tools and open datasets have widened the discussion on migration. However, challenges remain to focus the discussion by conducting research that can be directly reproduced and built on further by an interdisciplinary community. Although case studies provide a powerful lens to understand migration processes, there is value in being explicit and rethinking what case studies can have in common. This is important because starting any study from first principle is time consuming. Time pressures can often lead to opaque analysis that compromise the opportunity for the critical discussion they merit. In the long run, this leaves undesirable room for inconsistent results and incomparable analysis.

Being explicit about methods is especially important in interdisciplinary research, where stakeholders, by definition, come from different backgrounds and with different assumptions and conventions. Migration is one such topic as it has its place in the fields of Geography, Sociology, Economics and History to name a few. To understand patterns of migration and keep pace with the preceding research, the researcher is expected to be fluent in three distinct areas: the mathematics, computational implementation and contextual interpretation of their research. This is no inconsiderable task, as paying less attention in one could be detrimental to the other two. As larger datasets and better computation become the norm in research, there should be a shift in research towards to placing greater value on communication of the process beyond just simple code sharing. The authors argue that there is merit to the literate programing paradigm, specifically the 'recipes approach' of communicating methods.

* boyana.buyuklieva.14@ucl.ac.uk

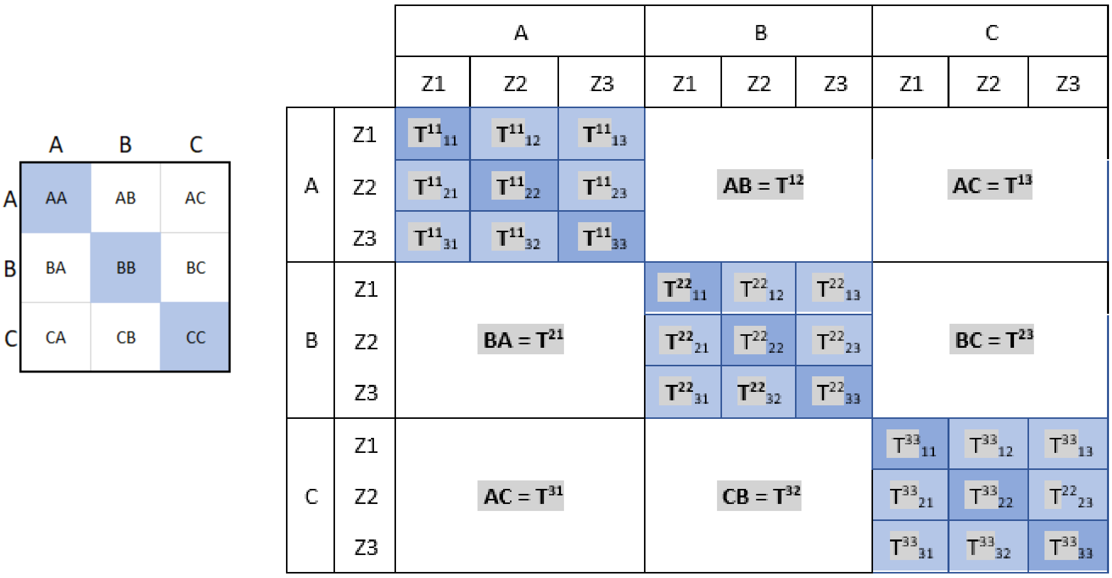
† a.dennett@ucl.ac.uk

1.1. Review

Uncertainty can creep into the study of migration at various points in the research process. For example, definitional issues are well known and have been discussed widely (Long, 1991, Rees, 1977). The common census question used to record migrants (Did you live in the present location on this date X years ago?) reveals an imbalance between the accuracy of the record kept for in-movers and out-movers. Indeed, it has been noted that:

> "The history of migration modelling is replete with contradictory theory and conflicting empirical evidence." (Fotheringham et al, 2004 p1640)

However, further uncertainty can emerge when attempting to analyse migration data. A discussion on data available for migration research in the general case is provided in Bell et al. (2015) and Long (1991). Stillwell et al. (2010) provide a higher-level overview of datasets specifically in UK. In general, however (and this includes most census-based sources), data relate to single transitions between one origin and one destination and thus are available in paired list or, frequently, matrix form (See figure 1).



	A	B	C
A	AA	AB	AC
B	BA	BB	BC
C	CA	CB	CC

		A			B			C		
		Z1	Z2	Z3	Z1	Z2	Z3	Z1	Z2	Z3
A	Z1	T^{11}_{11}	T^{11}_{12}	T^{11}_{13}	AB = T^{12}			AC = T^{13}		
	Z2	T^{11}_{21}	T^{11}_{22}	T^{11}_{23}						
	Z3	T^{11}_{31}	T^{11}_{32}	T^{11}_{33}						
B	Z1	BA = T^{21}			T^{22}_{11}	T^{22}_{12}	T^{22}_{13}	BC = T^{23}		
	Z2				T^{22}_{21}	T^{22}_{22}	T^{22}_{23}			
	Z3				T^{22}_{31}	T^{22}_{32}	T^{22}_{33}			
C	Z1	AC = T^{31}			CB = T^{32}			T^{33}_{11}	T^{33}_{12}	T^{33}_{13}
	Z2							T^{33}_{21}	T^{33}_{22}	T^{22}_{23}
	Z3							T^{33}_{31}	T^{33}_{32}	T^{33}_{33}

Figure 1: Origin and Destination Matrix with 3 locations (A, B and C - simplified on the left), and a more complex case with sub-locations based on Dennett and Wilson, 2011 on the right.

Depending on the source, OD matrixes such as those shown in Figure 1 can be incomplete: they might exclude within area moves (such as with migration data obtained from GP patient registers in the UK), only provide row or column totals (or even just net-flows if obtained from the residuals of population change) or otherwise have missing data due to geographic aggregation or other errors. These issues can present problems for any researchers looking to replicate work or build on previous findings, and although larger datasets and better computation make complex modelling very tempting, it is arguably now more important than ever to set solid foundations by creating a common understanding of baseline metrics that can be produced in accordance with whatever data a researcher has at hand.

Underpinning any piece of successful quantitative analysis are the basic descriptive statistics of that field. They are the foundation for more advanced models and as Plane (1982, p.441) notes, statistics like prior migration probabilities 'may be chosen simply to represent an observed base period matrix of flow or, alternatively, to represent an initial guess'. Therefore, care must be taken to evaluate these basic measures unambiguously, as miscalculation will only widen the margin of error in more complex information-theoretic models. A probability or rate of migration might be very different depending on the denominator. This would normally be the population 'at risk' of experiencing a migration event (for example, if looking at migrants aged 16-24, then we would only consider this age group in the population denominator), but is this the event of departure (origin population); arrival (destination population) or some combination of the two? The ability to model and estimate robustly is only as good as the researcher's ability to accurately pin down the basic patterns and trends.

Describing basic patterns and trends involves tapping into the research of others, instead of starting from first principles. The first step to avoid re-inventing the wheel is understanding what methods we can use for the foundation of our study, which in turn involves getting a clear understanding of whether the data one has at hand is similar or different to what other researchers have worked with.

In his seminal paper, Knuth (1984) presents a computer language implementation for what he calls 'literate programming'. Promoting Knuth's WEB language or any other specific tool is not the purpose of this paper, as the plethora of GIS applications all have their unique virtues. Instead, this paper will engage with the paradigm of literate programing:

> "Let us change our traditional attitude to the construction of programs: Instead of imagining that our main task is to instruct a computer what to do, let us concentrate rather on explaining to human beings what we want a computer to do." (Knuth,1984 p.97)

Knuthian literate programming is about writing code that reads like prose, so that meaningful details that pertain to the data processing, structure and computational implementation of formulas are explained, not assumed and omitted. In relation to this, what we identify as problematic for healthily migration research is the conceptualisation of different studies in terms of 1) data used and 2) metrics calculated. The first issue pertaining to data can be addressed through producing a visual data schema. The second problem relating to ambiguity in the migration metrics, generally relates to two main aspects:

1. Are moves within an area included?
2. When considering a 'population at risk (of experiencing a migration event)' for generating some migration rate: Do you use the population at the origin, that at the destination or both in the denominator?

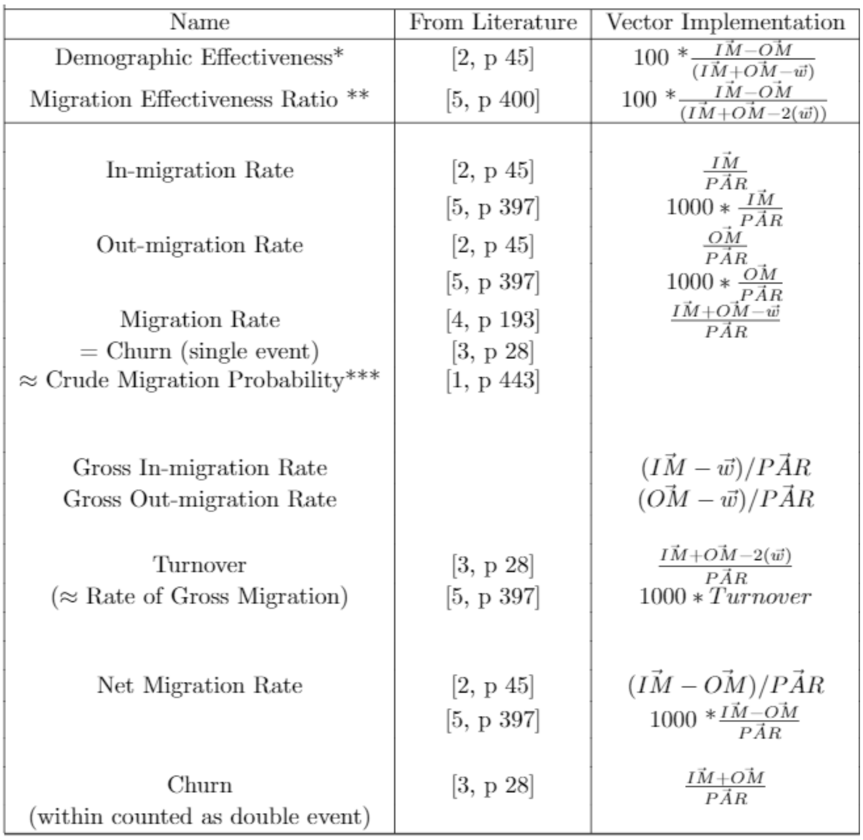
This problem can be addressed through transforming metrics from mathematical notation to pseudocode, which is expressed in simple, language-independent vector arithmetic that allows for no simple mistakes such as wrongly placed parenthesis.

These solutions are now exemplified below in a 'recipe' example:

2. Recipes for Migration Research - Ingredients and Preparation

2.1. Ingredients

The metrics in Table 1 (our completed dishes) are all comprised of widespread raw migration ‘ingredients’: origin/destination flows (total in, out and within area moves) and populations (at risk of experiencing flow events).



Name	From Literature	Vector Implementation
Demographic Effectiveness*	[2, p 45]	$100 * \frac{\vec{IM}-\vec{OM}}{(\vec{IM}+\vec{OM}-\vec{w})}$
Migration Effectiveness Ratio **	[5, p 400]	$100 * \frac{\vec{IM}-\vec{OM}}{(\vec{IM}+\vec{OM}-2(\vec{w}))}$
In-migration Rate	[2, p 45]	$\frac{\vec{IM}}{\vec{PAR}}$
	[5, p 397]	$1000 * \frac{\vec{IM}}{\vec{PAR}}$
Out-migration Rate	[2, p 45]	$\frac{\vec{OM}}{\vec{PAR}}$
	[5, p 397]	$1000 * \frac{\vec{OM}}{\vec{PAR}}$
Migration Rate	[4, p 193]	$\frac{\vec{IM}+\vec{OM}-\vec{w}}{\vec{PAR}}$
= Churn (single event)	[3, p 28]	
≈ Crude Migration Probability***	[1, p 443]	
Gross In-migration Rate		$(\vec{IM}-\vec{w})/\vec{PAR}$
Gross Out-migration Rate		$(\vec{OM}-\vec{w})/\vec{PAR}$
Turnover	[3, p 28]	$\frac{\vec{IM}+\vec{OM}-2(\vec{w})}{\vec{PAR}}$
(≈ Rate of Gross Migration)	[5, p 397]	$1000 * Turnover$
Net Migration Rate	[2, p 45]	$(\vec{IM}-\vec{OM})/\vec{PAR}$
	[5, p 397]	$1000 * \frac{\vec{IM}-\vec{OM}}{\vec{PAR}}$
Churn	[3, p 28]	$\frac{\vec{IM}+\vec{OM}}{\vec{PAR}}$
(within counted as double event)		

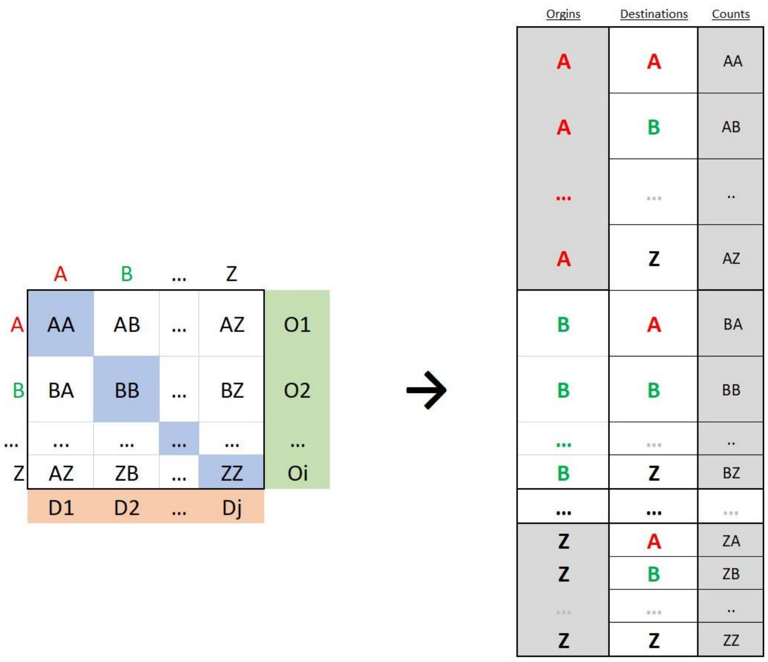
[1] Bell et al. (2002) **[2]** Boyle et al. (1998) **[3]** Dennett and Stillwell (2008) **[4]** Hinde (1998) **[5]** Rowland (2003)

Table 1: Basic vector implementations with exact reference where necessary

2.2. Preparation

<u>*Step 1:*</u> *Rearrange your matrix into a long (paired list) table.*

Converting the matrix into a long table creates a ‘tidy’ object which is easier to use (Wickham and Grolemund 2016, p 149). The long data format treats each individual migration count as an independent observation and provides it with its own row. The columns are then reserved for variables, which for now are the origin and destination of each count.

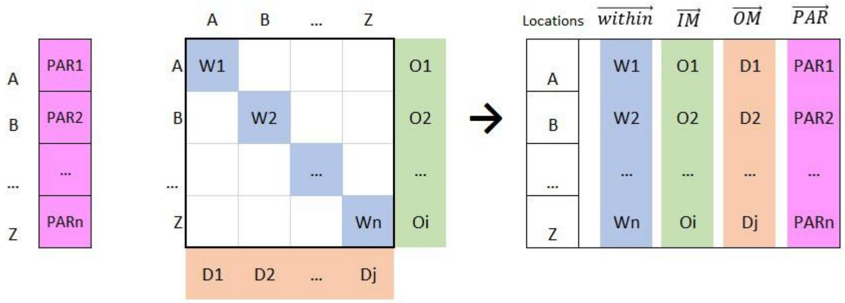


	A	B	...	Z	
A	AA	AB	...	AZ	O1
B	BA	BB	...	BZ	O2
...	...	...	...	...	...
Z	AZ	ZB	...	ZZ	Oi
	D1	D2	...	Dj	

Orgins	Destinations	Counts
A	A	AA
A	B	AB
...	...	..
A	Z	AZ
B	A	BA
B	B	BB
...	...	..
B	Z	BZ
...	...	...
Z	A	ZA
Z	B	ZB
...	...	..
Z	Z	ZZ

Figure 2: Step 1

Step 2: *Create your four Basic Summary Vectors.*
These represent: within zone migration, in-migration to, out-migration from and resident population at risk (of migrating within, to or from) that zone.



A	PAR1
B	PAR2
...	...
Z	PARn

	A	B	...	Z	
A	W1				O1
B		W2			O2
...			...		...
Z				Wn	Oi
	D1	D2	...	Dj	

Locations	$\overrightarrow{within}$	$\overrightarrow{IM}$	$\overrightarrow{OM}$	$\overrightarrow{PAR}$
A	W1	O1	D1	PAR1
B	W2	O2	D2	PAR2
...	...	...	...	...
Z	Wn	Oi	Dj	PARn

Figure 3: Step 2

Step 3: Match your long table and Basic Summary Vectors.
This can be done in one of two ways which are equivalent when calculating aggregate migration stocks, but different if individual migration counts between two locations are considered. For the latter, when the population at the orgin is used for the matching, then the interpretation is population at risk of moving away. For rate on migration counts where matching is done on the destination, the interpretation is population at risk of being a new arrival.

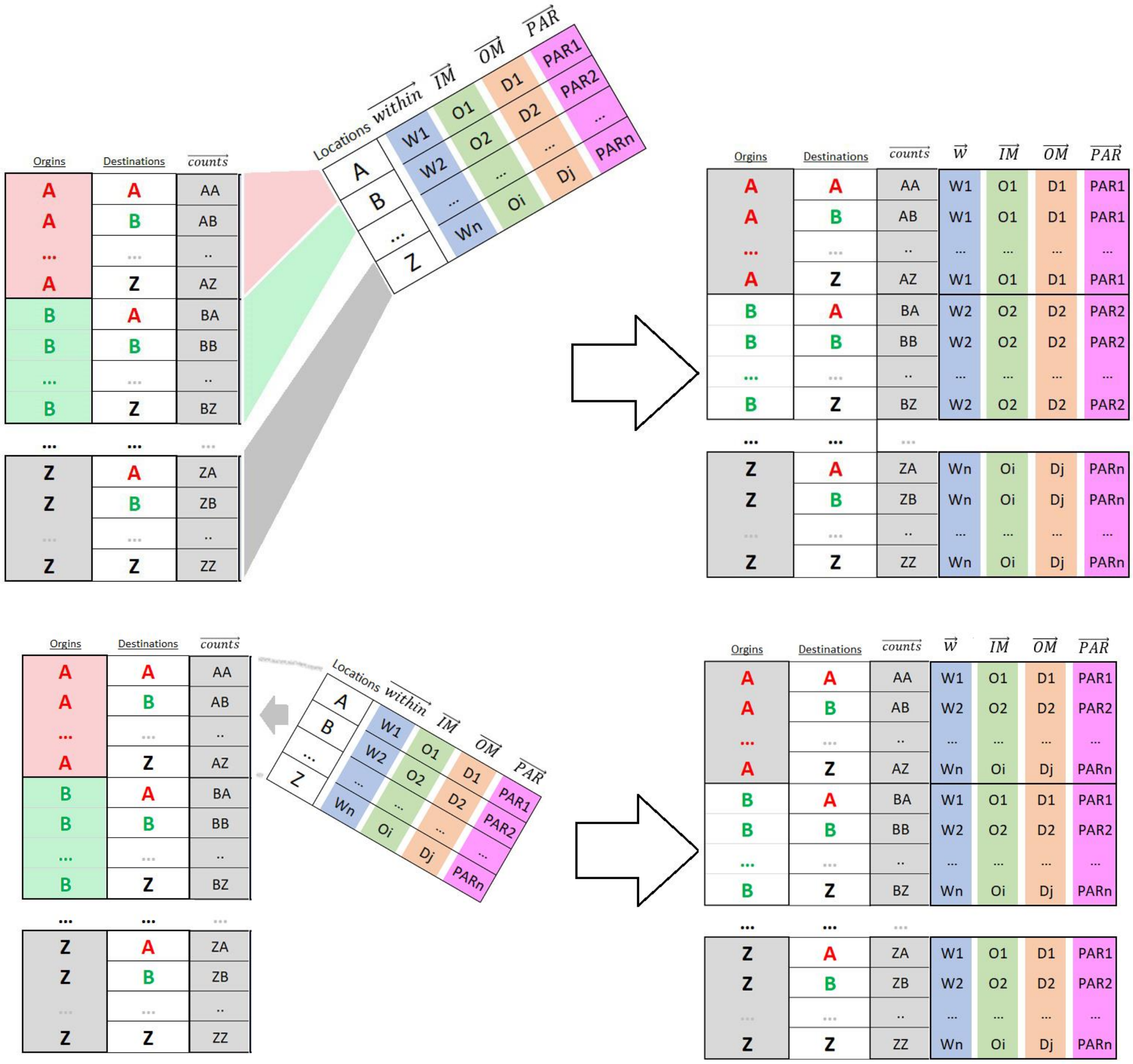
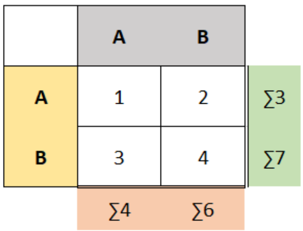


Figure 4: Step 3 (Above: Matching by Origin / Below: Matching by Destination)

<u>Step 3:</u> Plug the column vectors created in Step 2 into the metric of choice from Table 1.

Things can go wrong if the basic vectors are not considered as a block and matched consistently on an origin or a destination. For example, if the basic vectors from the matrix are match based on the origin, but the population at risk was matched on the destination this would give inconsistent results.

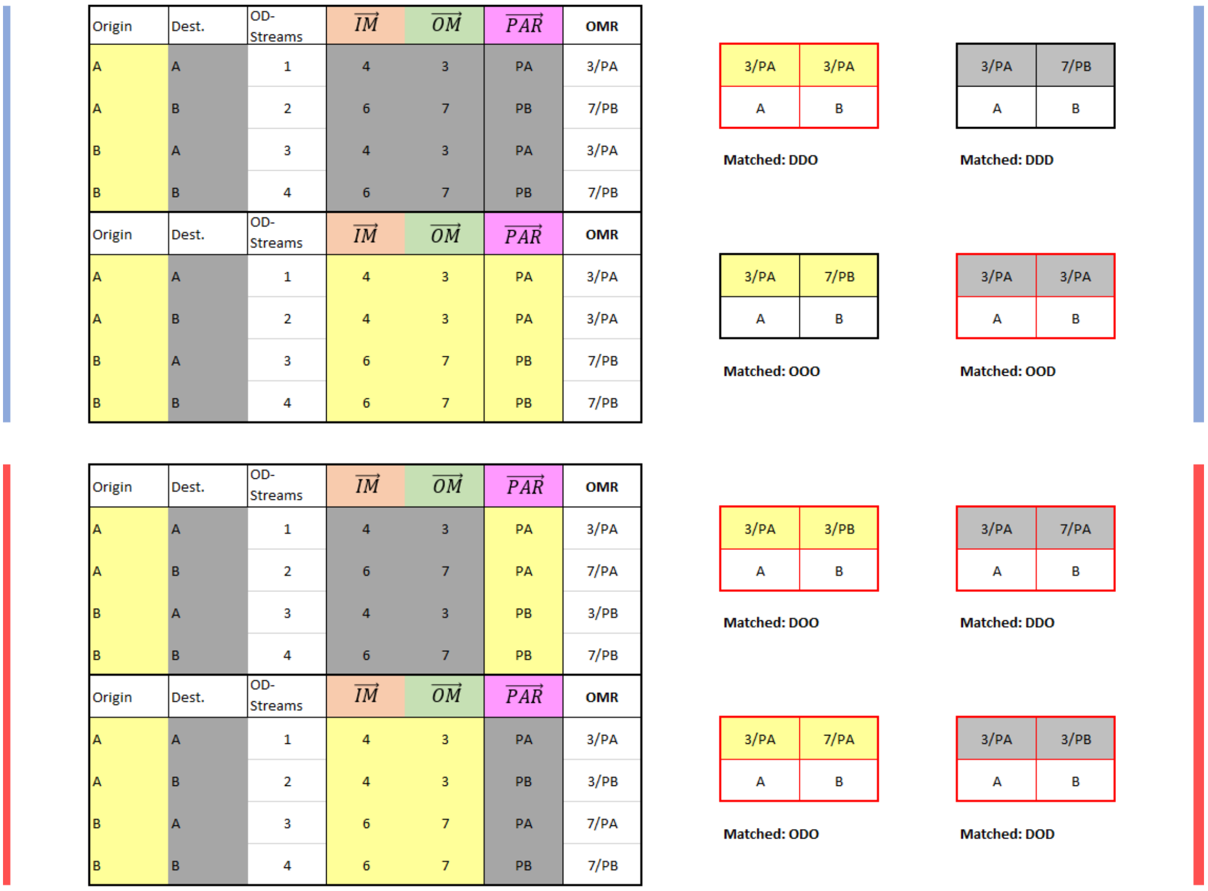
3. Getting the Recipe Wrong



	A	B	
A	1	2	∑3
B	3	4	∑7
	∑4	∑6	

Figure 5: Example 2x2 Matrix

Consider the matrix in Figure 5 with its row and column totals each side. Now consider that the population at risk needs to be matched to do some mathematical operation, say calculate rate of out-migration (OMR). The above task involves deciding which geometry to match to three times: once for the matrix summaries (i.e row and column totals), a second time for the resident population (PAR) and finally to visualise the results. The first two decisions lead to 4 possible 2-dimensional array combinations (Figure 6, left).



Origin	Dest.	OD-Streams	$\overrightarrow{IM}$	$\overrightarrow{OM}$	$\overrightarrow{PAR}$	OMR
A	A	1	4	3	PA	3/PA
A	B	2	6	7	PB	7/PB
B	A	3	4	3	PA	3/PA
B	B	4	6	7	PB	7/PB

3/PA	3/PA
A	B

Matched: DDO

3/PA	7/PB
A	B

Matched: DDD

Origin	Dest.	OD-Streams	$\overrightarrow{IM}$	$\overrightarrow{OM}$	$\overrightarrow{PAR}$	OMR
A	A	1	4	3	PA	3/PA
A	B	2	4	3	PA	3/PA
B	A	3	6	7	PB	7/PB
B	B	4	6	7	PB	7/PB

3/PA	7/PB
A	B

Matched: OOO

3/PA	3/PA
A	B

Matched: OOD

Origin	Dest.	OD-Streams	$\overrightarrow{IM}$	$\overrightarrow{OM}$	$\overrightarrow{PAR}$	OMR
A	A	1	4	3	PA	3/PA
A	B	2	6	7	PA	7/PA
B	A	3	4	3	PB	3/PB
B	B	4	6	7	PB	7/PB

3/PA	3/PB
A	B

Matched: DOO

3/PA	7/PA
A	B

Matched: DDO

Origin	Dest.	OD-Streams	$\overrightarrow{IM}$	$\overrightarrow{OM}$	$\overrightarrow{PAR}$	OMR
A	A	1	4	3	PA	3/PA
A	B	2	4	3	PB	3/PB
B	A	3	6	7	PA	7/PA
B	B	4	6	7	PB	7/PB

3/PA	7/PA
A	B

Matched: ODO

3/PA	3/PB
A	B

Matched: DOD

Figure 6: All possible ways to match the resident population and matrix summaries in a 2-dimensional array (left). All possible ways to color area polygons based on each array (right).

Imagine the cells on the right correspond to geographic area polygons, where each (A) and (B) are shapes to be coloured in the colour of the OMR column value (a choropleth map). Zones can be coloured based on the polygon of the origin or that of the destination, hence the yellow and grey versions. (Assume a computer will colour these, therefore if you see multiple instances of the polygon ID, you take the first location (A) you see and move on to the next (B).)

If the summary vectors are matched to zones as described in Figure 6 (highlighted in blue), two results will be produced no matter whether the match is achieved through joining on the destination or origin columns. One of the results will be that: (A) will be colored 3/PA and (B) will be coloured 7/PB. The other result will be that all the polygons are coloured the same way (in the colour 3PA) – an obvious mistake.

If the new PAR dataset is matched in the remaining two ways (highlighted in red) two possible results are also generated: where, (A) = 3/PA and (B) = 3/PB, or, (A) = 3/PA and (B) = 7/PA. This is a subtler mistake. Firstly, because it is correct for the first row but wrong for all the following rows. Secondly, because it will still produce a variety of results or colours. It is easy to begin interpretation as if the indices have been derived from fully combining the OD matrix and the PAR vector. However, results will either 1) All be operations using only the resident population at (A) (effectively using only 1 row of the original PAR vector) or 2) perform the operation on all the PARs, but only using the aggregate row sum totals for (A) (indices then only use one row of the original basic matrix vector).

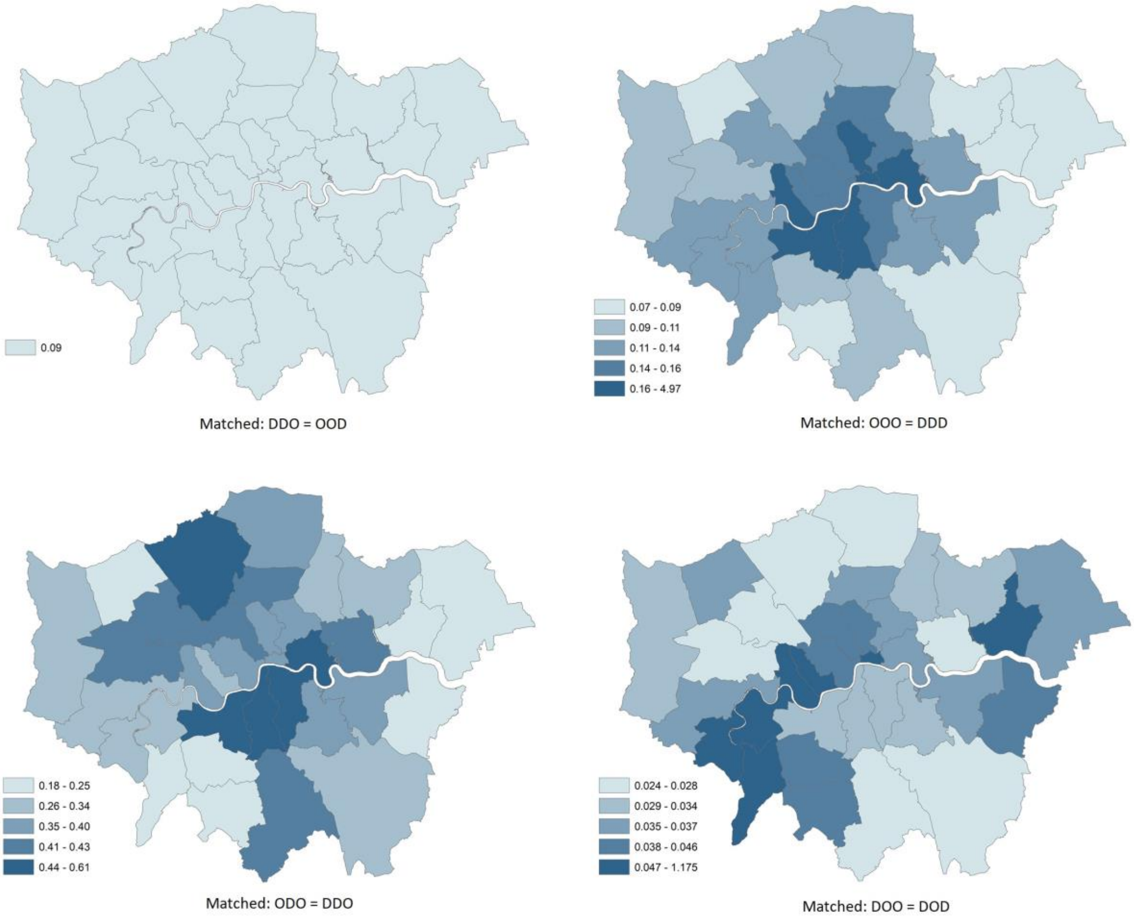


Figure 7: Choropleth maps plotted by 5 class quantiles with consistent matching (above) and inconsistent matching (below)

Figure 7 shows an example of matching and plotting using the consistent (blue method) and inconsistent (red method), where the basic vectors are not considered as a block. This shows that data structure is something classic mathematical notation cannot account for, but which is very important especially as dataset become larger and operations become harder to double-check by hand.

4. Summary and Conclusion

This work presented a recipe style workflow of migration metrics that is guided by the paradigm of literate programing. It raises the issue of reproducibility, specifically related to dearth access to migration data and ambiguous computational implementation and suggests that these can be remedied using simple diagrams. The main argument of the paper is that being clear by using visual aids is important because formulas on their own cannot help with data structure. The latter is important because not having a clear handle on your data can lead to subtle mistakes and further complications as datasets become bigger.

The authors argue that being explicit, especially in the interdisciplinary academic setting is paramount. This is the case as ‘blog - tutorials’ can be ephemeral (subject to link-rot) and lack peer-review. Academia addresses these points through an established system of record keeping. We hope that being explicit in the tradition of literate programing will offer recipes worth saving in the research of migration and other interdisciplinary fields.

5. Acknowledgements

This work was supported by the Economic and Social Research Council Engineering and Physical Sciences Research Council (EPSRC). Data provided for this study is from the Office for National Statistics 2011 Census. The exact tables used are *MM01CUK, Cell 1* for the OD Matrix and *KS101EW - Usual resident population* for the population at risk.

6. Biography

Boyana Buyuklieva (Bonnie) is a PhD student at the Bartlett Centre for Advanced Spatial Analysis, University College London. Bonnie is interested in residential mobility and its relation to housing and demographics at an urban scale.

Adam Dennett is a Senior Lecturer and Deputy Director of the Bartlett Centre for Advanced Spatial Analysis, University College London. Adam is a Population Geographer with interests in quantitative methods, spatial modelling and GIS.

References

Bell, M., Blake, M., Boyle, P., Duke-Williams, O., Rees, P., Stillwell, J., Hugo, G., 2002. Cross-national comparison of internal migration: issues and measures. Journal of the Royal Statistical Society: Series A (Statistics in Society) 165, 435–464.

Bell, M., Charles-Edwards, E., Kupiszewska, D., Kupiszewski, M., Stillwell, J., Zhu, Y., 2015. Internal Migration Data Around the World: Assessing Contemporary Practice: Internal Migration Data Around the World. Population, Space and Place 21, 1–17. https://doi.org/10.1002/psp.1848

Boyle, P.J., Halfacree, K., Robinson, V., 1998. Exploring contemporary migration. Longman, Harlow.

Dennett, A., Wilson, A., 2013. A Multilevel Spatial Interaction Modelling Framework for Estimating Interregional Migration in Europe. Environment and Planning A 45, 1491–1507. https://doi.org/10.1068/a45398

Fotheringham, A.S., Rees, P., Champion, T., Kalogirou, S., Tremayne, A.R., 2004. The Development of a Migration Model for England and Wales: Overview and Modelling Out-Migration. Environment and Planning A 36, 1633–1672. https://doi.org/10.1068/a36136

Hinde, A., 1998. Demographic methods. Arnold, London ; New York.

Holland, S.C., Plane, D.A., 2001. Methods of Mapping Migration Flow Patterns. Southeastern Geographer 41, 89–104. https://doi.org/10.1353/sgo.2001.0016

Knuth, D.E., Knuth, D.E., 1986. TEX: the program, Computers & typesetting. Addison-Wesley Pub. Co, Reading, Mass.

Long, L., 1991. Residential Mobility Differences among Developed Countries. International Regional Science Review, 14(2), pp.133-147.

Plane, D.A., 1982. An information theoretic approach to the estimation of migration flows. Journal of Regional Science 22, 441–456.

Rees, P., Bell, M., Kupiszewski, M., Kupiszewska, D., Ueffing, P., Bernard, A., Charles-Edwards, E., Stillwell, J., 2017. The Impact of Internal Migration on Population Redistribution: an International Comparison: The Impact of Internal Migration. Population, Space and Place 23, e2036. https://doi.org/10.1002/psp.2036

Rees, P.H., 1977. The measurement of migration, from census data and other sources. Environment and Planning A 9, 247–272.

Rowland, D.T., 2003. Demographic methods and concepts. Oxford University Press, Oxford ; New York.

Sjaastad, L.A., 1962. The costs and returns of human migration. Journal of political Economy 70, 80–93.

Dennett, A., Stillwell, J., 2008. Population turnover and churn: enhancing understanding of internal migration in Britain through measures of stability. Popul Trends 24–41.

Stillwell, J., Duke-Williams, O. and Dennett, A. (eds.) (2010) Technologies for Migration and Commuting Analysis: Spatial Interaction Data Applications. IGI Global, Hershey

Thomas, D.S.T., 1941. Social and economic aspects of Swedish population movements, 1750-1933,. New York,.

Wickham, H., Grolemund, G., 2016. R for data science: import, tidy, transform, visualize, and model data, First edition. ed. O'Reilly, Sebastopol, CA.

Williams, R., 1999. 1.1 What Is Literate Programming?. [online] Ross.net. Available at: http://www.ross.net/funnelweb/tutorial/intro_what.html [Accessed 12 Nov. 2018].