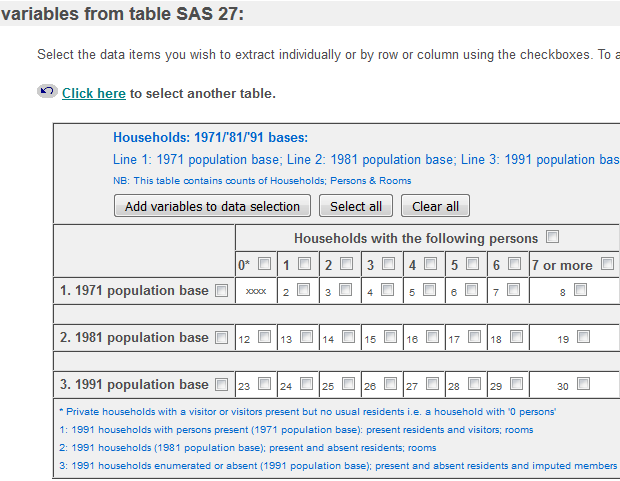
# Meen Census Stitching diary 2016

# Random things to follow up

..

|  |  |
| --- | --- |
| **Thing** | **Notes about thing** |
| Overall change in population | The stock approach to looking and CoB groups change - how to set that against the **overall** stockchangehappening as population increases? Absolute vs % also important here. Hmm.  Cf. Walford/Hayles 2012 p.296: different censuses had different ways of counting total population. It'd be useful to know how much difference this could make.  There was a Dorling paper about this as well, wasn't there?  Dorling, D., and D.J. Atkins. ‘Population Density, Change and Concentration in Great Britain 1971, 1981 and 1991’.  Which mentions this for defs:  <https://census.ukdataservice.ac.uk/media/51162/1991_defs.pdf> |
| "A number of variations exist between questions asked in Northern Ireland and Scotland compared with those of England and Wales (ONS 2004a)"  In Walford/Hayles 2012 p.296 | Will need to check on this in case it affects stuff. If initially looking at Scotland, less of a problem hopefully. |
| 1971 data snag: 111 more obs in CoB data than there are zones in the boundary data. | Out of ~15k. Leaving out for now but will need to note. |
| ~~1971 ED zones: boundaries over-run known geog edges~~ | ~~So I could subtract those to get a little more accuracy - although in the places where this happens, mostly it won't make much difference (certainly relative to voronoi-ness)~~  ~~Without doing this (or another fix e.g. fractions of the sums of areas from the intersect) the numbers won't sum properly.~~  ~~Actually, that other fix ends up being precisely the same as doing it first, so I'm just doing that!~~ |

Note this. 1991 SAS -> what's with the different population bases? Quick look at the data, the numbers are pretty radically different and apparently in no sort of consistent way. What's the crack?



## Some bits and bobs

Vertovec, Steven. ‘Super-Diversity and Its Implications’. Ethnic and Racial Studies 30, no. 6 (1 November 2007): 1024–54. doi:10.1080/01419870701599465.

<http://www.tandfonline.com/doi/abs/10.1080/01419870701599465#.VpU_v1Jy0bM>

A very nice overview of UK CoB migration. Via superdiversity conference. Quite old now, no? But many many citations so doubtless Gwilym already knows about it...

## Useful links

Ethnicity vs CoB: <https://www.ucl.ac.uk/celsius/online-training/ethnicity/et040200>

<https://www.ukdataservice.ac.uk/get-data/themes/ethnicity>

How LA geographies change: <https://www.nomisweb.co.uk/forum/posts.aspx?tID=478>

Datazone matching guide: <http://www.gov.scot/Resource/0048/00483471.pdf>

Some "population by nationality and CoB" links: <http://webarchive.nationalarchives.gov.uk/20160105160709/http://ons.gov.uk/ons/taxonomy/index.html?nscl=population+by+nationality+and+country+of+birth>

Guide to 1991 and earlier censuses: <https://www.ons.gov.uk/census/2001censusandearlier/aquickguideto1991andearliercensuses#census-geography-1801-2001>

..

Census definitions n ting:

<https://census.ukdataservice.ac.uk/media/51156/1971_defs.pdf>

<https://census.ukdataservice.ac.uk/media/51162/1991_defs.pdf>

# Census geog hierarchies at different times

## 1991 England

* EDs / Wards / Districts / Counties.

## 2011 England

* OA / LSOA / MSOA / Local authority / country

# Journo

So pulling together stuff after meeting Geoff Meen. First job is all just very data. But let's make some notes on that first.

Other key note:

[Stitching census data](https://www.evernote.com/shard/s312/nl/43551379/8578db20-16eb-48e1-b4d5-d392bba43b23)

..

Things from the meeting:

<https://www.essex.ac.uk/history/research/icem/>

I-cem: is only for 1851-1911. Which might come in handy I guess.

My first job:

* 71-2011 data
* For one city to start with (Glasgow) to test
* Then whole of UK. Focusing on cities, prob.
  + TTWAs would seem sensible but LAs probably don't nest into those, do they?
  + Naw - wards nest into them.
  + p.s. <https://en.wikipedia.org/wiki/State_of_the_English_Cities> "State of the English Cities"
    - Ongoing... or not. I have a feeling it isn't. But it was very urban-indicatorsish, wasn't it?
      * <http://webarchive.nationalarchives.gov.uk/20060731065549/communities.gov.uk/index.asp%3Fid%3D1127498>
  + Oh actually, yes, they might nest (wards and LAs do, I think).

But they don't match with census boundaries, do they? Oh. yeah, maybe:

<http://www.nisra.gov.uk/archive/census/2011/note-on-statistical-and-administrative-geographic-units.pdf>

"Small Areas (statistical unit) nest within Super Output Areas (statistical unit)which in turn nest within electoral wards (administrative unit)."

Oh and Scotland/NI being their own orgs too. So.

<https://census.ukdataservice.ac.uk/use-data/guides/boundary-data>

Ah. Country / LA / MSOA/ LSOA / Output area. But something different for Scotland, presumably (though data comes from the same source).

<https://www.nomisweb.co.uk/forum/posts.aspx?tID=478>

So: LA boundaries changed. Well! And what's the equivalent of LA boundaries in Scotland...? Err. They appear to be Local Authorities too.

<http://saspac.org/support/1991-census/>

"There are two predefined sets of cross-tabulations produced from the 1991 Census: Small Area Statistics (SAS) and Local Base Statistics (LBS). There are 86 tables in the SAS dataset and 99 in the LBS. The SAS and LBS tables are similar in structure but SAS are available for the smallest geographical level from the Census (Enumeration District), whereas the LBS are only available ward geographies and above."

1971 and 1981 start at SAS / go down to enumeration district level. It's ED / Ward / District / County / Country. Do we have ED boundaries available?

## GEOGRAPHEEEEZ

Ah, OK. Well, let's start here then:

<http://ons.gov.uk/ons/guide-method/geography/beginner-s-guide/index.html>

* "Census data can be produced for most geographies by best-fitting from OA to the required output geography current on 31 December 2011."
* Output area, smallest. Super output area next up.

OK, just going through the infuse page to see what geographies they've got. "Currently on country-of-birth (detailed)"

..

<http://www.ons.gov.uk/ons/guide-method/geography/products/census/lookup/index.html>

"2011 Census estimates for all geographies (except national parks) are best-fitted from output areas (OA) in line with the Geography Policy for National Statistics."

Oh ah. OK, so that's where the LA data comes from. Not necessarily nested. There's a guide to how that's done there as well. (Though the link is broken on that page).

So: all built on the basic OAs. Except, oh - look below, LA districts fit perfectly. So that's nice.

<http://www.ons.gov.uk/ons/external-links/social-media/g-m/2011-oas-to-2011-lower-layer-super-output-areas--lsoas---middle-layer-super-output-areas--msoa--and-lads.html>

### AH HA! (from searching "census boundary change 1971 to")

<http://www.ons.gov.uk/ons/guide-method/census/1991-and-earlier-censuses/guide-to-earlier-census-data/census-geography-1801-2001/index.html>

So:

* 1974: major boundary re-organisation. So 71-81 will take some adjustment.
* England/Wales unitary authorities: 94-97. So change between 91 and 01 census.
  + Then "turnover" of other geog units - altering of various boundaries. But which?
* "The large majority of census results below the national level have been presented for areas of local government. From 1881 to 1971 these reflected an urban/rural divide, with Rural Districts and with various administrative types reflecting the size and status of urban areas."
  + So LAs are along that line.
  + But note 1971: urban/rural divide built into those. Wonder what they'll look like?
* London was reorganised in 1966 **then** the rest of the country in 1974.

..........

### MORE AH-HA!

<http://sec.kingston.ac.uk/research/research-centres/centre-for-earth-and-environmental-science/projects/censusbased/>

So there's the PDF presentation (at GISRUK!) presented there. And there's a paper cited there I'm just looking at, though the presentation looks more up to date.

Walford, Nigel. ‘Connecting Historical and Contemporary Small-Area Geography in Britain: The Creation of Digital Boundary Data for 1971 and 1981 Census’. International Journal of Geographical Information Science 19, no. 7 (1 August 2005): 749–67. doi:10.1080/13658810500071832.

And chasing citations for that:

* Ah: <http://www.tandfonline.com/doi/abs/10.1080/17445647.2014.960484?journalCode=tjom20>
  + That is Stephen Clark... who may or may not still be at Leeds.
* Then Walford again:
  + Walford, Nigel. ‘Development and Design of a Web-Based Interface to Address Geographical Incompatibility in Spatial Units’. Environment and Planning A 45, no. 7 (July 2013): 1713–33. doi:10.1068/a45377.
  + From that: it's not currently publically available but... I wonder if I asked for specific use. I'm not even sure what years it works across...
    - 81/91/01.
* And via Walford's "Thirty Years of Geographical (In)consistency in the British Population Census", we *of course* find:
* Norman, Paul, Philip Rees, and Paul Boyle. ‘Achieving Data Compatibility over Space and Time: Creating Consistent Geographical Zones’. International Journal of Population Geography 9, no. 5 (1 September 2003): 365–86. doi:10.1002/ijpg.294.

For completion's sake, for those as cite Walford's 'connecting historical' there's:

* Norman, Paul, and Mylene Riva. ‘Population Health across Space and Time: The Geographical Harmonisation of the Office for National Statistics Longitudinal Study for England and Wales’. Population Space and Place 18, no. 5 (October 2012): 483–502. doi:10.1002/psp.1705.
* Which is 2012 and has a **different method** than the others, I think.

Norman, Paul. ‘Identifying Change Over Time in Small Area Socio-Economic Deprivation’. Applied Spatial Analysis and Policy 3:

* "In the UK, the geography used for the census is largely aligned with administrative geography so that data are published for the hierarchy of nation, regions and local government districts, often referred to as local authorities (LAs)."
* Oh OK, that clears that up. So when did LA boundaries change and how can we interpolate between them?
* Census Area Statistics: wards in England, CAS 'sectors' in Scotland. [113]

**And what, specifically, of Scottish geography?**

<http://www.gov.scot/Publications/2005/02/20697/52626>

OK, just looking at the download options from CASWEB. Interesting little factoid:

* **After 1981, Scottish data is always at Output Area leveL.** England/Wales are ED until 2001.
  + Though Reminder (via email from data service): 1981 Scots boundary data is marked as ED. It does, however, match the census OA zone IDs and they say it's the correct match.

## 2011 DOWNLOADING FROM INFUSE

Silly website: have to tick each specific category of CoB you might want. Fair enough! Least I've worked that out. Getting for all LAs.

Note:

* The only available unit category is "persons". Err, which kind of makes sense I guess.
* I've selected all categories, even though some are going to be sums of others. Should be obvious which those are, but might as well get it all.

OK, got, I think. Problem:

* No scotland. LAs included but not the actual numbers. So gonna need that from somewhere else.

Also just got:

* Cross-tab of country-of-birth vs year-of-arrival.

## 1971/1981 DOWNLOADING, CASWEB

Mainly because I want to see what kinds of geography we've got going on. I'm now putting together the general picture in a Word doc table instead, where hopefully I can start pulling that together into one sensible overview. But yes - it's going to be affected by what data I can actually get hold of. Let's take a look at 1971.

1971 and 1981 start at SAS / go down to enumeration district level. It's ED / Ward / District / County / Country. Do we have ED boundaries available?

Update:

* England/Wales were EDs until 2001, when OAs became the base unit. Scotland started with OAs in 1981.
* But currently:
  + There are only shapefiles for Scottish EDs in 1981.
  + 1971 EDs are "synthetic" - voronoi polygons or similar. Very crude at the city level e.g. Glasgow.
* Note: Gavin's been using datazones in Scotland and LSOAs in England. Equivalent apparently. Were datazones the "thing between OA and district"?

Just got this from the UK Data Service:

"I checked with a colleague who is more knowledgeable about census boundaries and I can confirm that the boundaries in UK Data Service Census Support that you can download from the Boundary Data Selection tool labelled as Enumeration Districts for 1981 are the correct boundaries that were used for the collection and dissemination of the 1981 census data. So they do correspond to the 1981 census data.

"One thing to be aware of, you may notice you have more areas in your census data then are available in the boundary data. This is because of things such as 'shipping areas' (i.e. people counted as being on ships at the 1981 census night in Scotland) where no boundary exists."

Good tip! So let's check - IDs should match up if this true. Or \*something\* should match up.

* 1991 Scots OAs: they do mostly match 1981 EDs, yes (though with obvious changes). Though 1981: much cruder boundaries.
* Ah yes: so:
  + Despite the **CASWEB census 1981 scotland download saying "OA"**, and the **edina borders data is ED for that year**, they do match IDs. They're output areas.
* And 1971: **IDs do appear to match also for census EDs and polygons.** Or there's one at least.
  + So there's some scope for aggregration to larger zones at the very least from those synth polygons.

## QUICK RECAP ON THE ACTUAL AIM

Looking back at Geoff's paper / the econometric equation: predict migrant numbers of CoB i in zone j at time t.

My current thoughts on the geographies:

* It might even be easier to look at local authority level for the UK as a whole - though the migrant numbers may be so low (per country group) that it's a problem. Though there'd be more than at a lower geography! (What's the 'right' number?)

## WICID / FLOW DATA

This link may not work without logging in first:

<https://wicid.ukdataservice.ac.uk/cider/wicid/query/select/data/tables_list.php?mode=showtab&amp;table=1&amp;dataset=56&amp;refper=5>

But a point:

* Breaking down by "born in UK / not in UK", it should be possible to test whether **internal** migration patterns are significantly different for the two groups.
  + Can probably do some aggregating to test this if numbers are not large.
* 2001 only has those two categories (UK/not UK born)
* Except... there will be **no way to control for socio-econ/demographic reasons**. Cf. Finney/Simpson "Internal migration and ethnic groups: evidence for Britain from the 2001 Census"

Lots of other potentially useful things. E.g.

* "Location of usual residence and place of work by country of birth" [2011 only]
  + Could be useful for checking on diffs in work location patterns, if any. For stuff on interaction of migrants/employment

## MORE ON 1971- 2011 GEOGRAPHIES

Various little bits. E.g.:

* From Dorling/Atkins 1995: 1971 EDs had grid references. (Which is presumably where the synth polygons come from). So **where can I get that original data?** Points would be better to work with.
  + "A changing definition of the population has been adopted in this study because it appears to produce trends which are consistent with the district level trends indicated by the official mid-year estimates (which are based on a constant definition of the population). For more general purposes of comparing census data between 1971 and 1991 these population definitions may not be appropriate."
  + Huuurm.
* "The basic areal units chosen for this study were 1981 Census wards because these areas held administrative significance at the midpoint of the study period, and because they were the smallest areas for which digital boundaries were available for research at the time of writing. In Scotland the areas to which Census wards refer are Part-Postcode Sectors which were adopted here because the same areas have been used in both the 1981 and 1991 Censuses. Most wards in use today are identical to those which were used to report statistics from the 1981 Census, and many wards have had unchanged boundaries since 1971." [3]

## NOT!!! TABLE OF GEOGRAPHIES

So for each census year, for England+Wales and Scotland separately (might just not think about NI for a little bit!), I could do with a table of geographies / how they've changed / what kind of boundary data, if any, is available.

...

OK, I'm over-complicating this. It's simple enough:

* Enumeration district / Ward / LA for everywhere.
  + For England until 2001 when OAs begin.
  + For Scotland only in 1971. 1981 is output areas - though it's 'enumeration district' in the boundary data (possibly because those two were the same geography). At any rate, they have the correctly matching ID codes. (Or there was one at least!)
* So from 1981 for Scotland / 2001 for England+Wales, Output Areas.
* For 2001 and 2011, LSOA and MSOA are available for England and Wales. But I can't see Data / Intermediate Zones for Scotland - might be available via Scottish census download.

Ah, and this is another thing to note, via:

<http://infuse.mimas.ac.uk/help/definitions/2011geographies/>

* **For the smallest areas, disclosure risk might mean CoB data has been randomised or can't be disclosed.**

## SCOTLAND CENSUS DOWNLOAD...?

<http://www.scotlandscensus.gov.uk/census-results>

Yup. Everything. I've just grabbed datazones for 2011 from "bulk download". Much less arsing about than inFuse, huh? Is there an equivalent for England? Probably not.

Note: there's randomisation going on at the datazone level, possibly. It doesn't say when the rules are applied so tricky to know.

Also - guide to Datazones (because I can't seem to find 2011 datazone boundaries via Edina. 2001 are there.)

"Data Zone 2001 and2011 Matching–Quick Guide": <http://www.gov.scot/Resource/0048/00483471.pdf>

..

* So in theory they don't change much. 500-1000 people.
* But there was a **massive boundary change** between 01 and 11 - about half of them are different.
  + Oh no - even less with "direct boundary match". The half figure is for population via postcode.
* That has the datazone ID coding convention: it's consecutive for 01 and 11, so 11 follows on from where 01 left off:
  + "Data Zones 2001 are labelled fromS01000001 to S01006505 (6,505 individual data zones) and the Data Zones 2011 are labelled S01006506 to S01013482 (6,976 individual datazones)".

Via here:

<http://saspac.org/data/2011-census/scotland-2011/>

... we seem to be able to get the boundaries. Let's look. Newp, those links broken - but the Scots website is looking promising:

<http://www.gro-scotland.gov.uk/statistics-and-data/geography/our-products/census-datasets>

Oh Lord, what's this now?? [http://www.gro-scotland.gov.uk/files//geography/products/1991-census-bkgrd.pdf](http://www.gro-scotland.gov.uk/files/geography/products/1991-census-bkgrd.pdf)

* "1971 SAS was retabulated for 1981 EDs, thus providing a degree of continuity over 3 censuses."
  + So does that mean the 1971 ED data download matches those 1981 borders...?
* Aaand a load of other detail making clear it's really bloody messy. Oh good!

...

<http://www.gov.scot/Publications/2005/02/20732/53083>

That has another couple of useful links:

"Small Area Statistics reference material": <http://www.gov.scot/Topics/Statistics/sns/SNSRef>

And this should, in theory, have datazone boundaries: <http://www.sns.gov.uk/>

Here? Yes, I think so!

<http://sedsh127.sedsh.gov.uk/Atom_data/ScotGov/StatisticalUnits/SG_StatisticalUnits.atom.en.xml>

Finally! And, yes, can clearly see the changes. So this might give us something to start with at least. Prior to 2001, not sure what intermediate geographies there might be.

## FLOW DATARRRAARR

Just reading this on 2011 flow stuff too (the webinar I missed): <https://www.ukdataservice.ac.uk/media/305012/CensusFlowDec2014Webinar.pdf>

* There **must** be something useful to be done with the "where were you one year ago" question to think about spatial persistence. But I'm not sure I can get cross-tables for CoB vs this question (vs background of how much other groups move). We'll see I guess!
* The "what is the address of your workplace" question may come in handy too, employment-wise.
* p.9/10: All the variables we're interested in: **they're going to be secure** - and it sounds like a VML solution hasn't been set up yet.
* This presentation's example is really useful:
  + **Foreign origin (i.e if in a different country the previous year**) can be mapped very accurately. May come in useful.

But for now, not a great deal of use...

## Changing questions 1971 to 2011

Update: I just shifted all of this to its own note, as it's got a bunch of CoB table screenshots in that I don't want here. Ref:

[Census 1971-2011: changing country of birth outputs](https://www.evernote.com/shard/s312/nl/43551379/0f94b21e-da9d-40a2-b836-da68636cb979)

Oh look! -> <http://infuse.mimas.ac.uk/help/definitions/2011geographies/index.html>

Contains all 2011 boundaries for download.

Note also: <http://www.ons.gov.uk/ons/guide-method/census/2011/the-2011-census/processing-the-information/statistical-methodology/statistical-disclosure-control-for-2011-census.pdf>

p.14 seems particularly useful: "No persons or data items are removed from the census data and therefore outputs at national level and high geographies are unaffected by record swapping. At all geographical levels table totals reflect the sum of the cells and there are no differences between cell counts in different tables. Swapped households have been matched on basic characteristics to preserve data quality, and most swapping has been done within the same local authority or middle layer super output area (MSOA). The level of non-response and imputation will actually have a far greater effect on any counts seen in the tables than record swapping."

The basic point: there's record swapping followed by randomised up-down adjustment. The proportions used are secret to avoid reverse engineering, though I'm not sure how you'd manage that.

## A LITTLE MORE ON GEOGRAPHIES...

Back from thinking about matching CoB cats against each other.

I'm not sure what impact the scale will have and e.g. whether we're able to use 1971 at all given all those boundaries are voronoi polygons.

Just trying to track down what the status of Census Tracts is - Walford uses them in the 30 yrs of inconsistency paper. See also:

<https://en.wikipedia.org/wiki/Census_tract#United_Kingdom>

Walford/30yrs: "These gaps have recently been filled on a nationally consistent basis by digitisation of the 1981 EDs from scanned images of the annotated census maps (Walford, 2005), and these were used to form digital boundaries for the 1971 –1981 CTs."

The CTs are not on Edina - might be a Walford-only product. Need to do a little reading on that. The 2005 paper...? Walford, Nigel. ‘Connecting Historical and Contemporary Small-Area Geography in Britain: The Creation of Digital Boundary Data for 1971 and 1981 Census’. International Journal of Geographical Information Science 19, no. 7

Quick scan of it...

And side-note:

<https://www.ucl.ac.uk/celsius/online-training/geography/ge030000>

I'm wondering about using districts - just because more of them might have kept consistent names/positions from 1971 and might not require any kind of actual boundary mapping at all. What does the next page from that UCL site say? (Note, this is for the Longitudinal Study, but it's the same geographies...)

<https://www.ucl.ac.uk/celsius/online-training/geography/ge030100>

* "For most of the timescale of the LS, the administrative boundaries in use are those defined in 1974 as a result of the 1972 Local Government Act. This replaced the pre-1974 geography of standard regions, counties and local authority districts (i.e. county districts and county boroughs) with a new geography with redrawn borders, and introduced new ‘metropolitan counties’ such as Greater London and new county district-level entities such as London boroughs, metropolitan districts and non-metropolitan districts."

Which I knew, didn't I? So there's a large change from 71 to 81. I think it might be worth concentrating on 81-11 for now. That gives us four censuses to work with and side-steps the (large) 71 boundary issues.

Note, that page is all about LS boundaries. They made some specific choices that I think don't apply to the census - e.g. there's been re-coding to post-74 district boundaries. It's still the case that 81-11 is probably the way to go for now. Let's print this and churn...

Printed above

Some links I seem to have not pasted in I'm re-getting from firefox history:

A bunch of great stuff on Scots geographies (though it doesn't mention intermediate zones):

<http://www.scotlandscensus.gov.uk/documents/supporting_information/2011_Census_Geographies.pdf>

Just has Scots LA boundaries in:

www.lgbc-scotland.gov.uk/includes/downloadfile.asp?file=/maps/LAs\_Scotland\_1995\_on.pdf

...

Well, having gone through the print-out above, I'm not sure I feel superbly more knowledgeable.

## ENGLAND OH ENGLAND

I'm confusing myself. Let's break this down by sticking first to:

* England
* 81 to 2011
* Maybe aggregating EDs/OAs up to 2011 ward boundaries.

Let's just look at those geographies first. Let's compare 2011 output areas / 1981 EDs and 2011 wards.

Much better approach! Something I can try: aggregating up to one consistent geography. Currently looking at 2011 ward boundaries - might be an idea to code this up to work with all options down the road. Possibly even with geotools, since that might end up being useful for the server. Though also might be overkill. (It should really have those kinds of functions, though, shouldn't it?)

## ENGLAND: AGGREGATING POLYGONS...

OK, so let's have a go at this. Testing with 2011 ward boundaries, which I now have. Was mulling geotools but I could do with getting this done more quickly. Though... for the scale of the job, also mulling python. E.g. look at geopandas, it's got just the kind of intersection stuff that might work -

<http://geopandas.org/user.html#geopandas-functions>

Bleurgh - not having a great deal of luck with python. Same fuckup of dependencies as with Jekyll. Not helpful.

This looks like a route to try with R instead:

<http://stackoverflow.com/questions/9977074/how-to-calculate-the-area-of-polygon-overlap-in-r>

Let's give that a go. Seems like there should be something in the rgeos package that would do the job though...

<http://gis.stackexchange.com/questions/52817/remove-intersect-of-one-polygon-from-another-in-r>

From the manual, gIntersection looks like it would do the job. I can't help but feel like it might need something similar to a hashgrid or something else to avoid doing a full test on every single pair. Perhaps it's already efficient enough, who knows? I need to try with a subset.

OK, so imagine that we can **get** the polygon intersections. What do we do with them then? The point isn't just to get the polygons themselves - I need to know where I'm redistributing values in those polygons from. Let's just look at what gIntersect does...

I'm now thinking of swinging back to geotools again as a source I can better control, particularly when it comes to the time taken to process.

Let's just try a smaller group of polys too - look, Robin!

<http://robinlovelace.net/r/2014/07/29/clipping-with-r.html>

Actually, that's not quite what I'm after - just want a set of complete polygons within a bounding box.

Well, see the code. Oh, I should tell you where that is:

**C:\Users\SMI2\Dropbox\SheffieldMethodsInstitute\Migration\MeenMigrationStitching\intersection.R**

---

So I'm definitely on R with this at the moment. Just about to test... hmm. Various interesting problems! Note this:

<http://gis.stackexchange.com/questions/35060/rgeos-drops-associated-values-when-intersecting-polygons>

The issue of getting the data. That link also mentions QGIS, which I'm just trying, but this is all still rather murky.

---

### QGIS!

Vector/geoprocessing/intersect - it does the intersect and then provides a merge of the data - so you get exactly the two zone IDs you need to do the area assignment. And I also think it doesn't take too long to run. Might be worth mulling how to then do more checking using Python - as I'll want to be doing various other sorts of checking along the way. Though... well, we'll see.

OK, well... let's run for Wards11 and EDs81. Then we can see about what to do with them. Note: the various slivers - they're small enough to not make much of a difference but it's probably a good idea to exclude those below a certain % of the area of the parent polygon. (Hence coding being a better idea than not - though I have a feeling the native geoprocessing tools in QGIS are in C and much faster.)

Though I've just noticed, I need to get Wales as well. Oops. OK, let's just get the wales data, shall we?

## Census geographies: restate assumptions

OK, so now I need to actually get this done. There's a few things I want to think about before getting on with it. It looks like I'll be doing the intersection stuff in QGIS - that gives me the merged ID data I need. But I'd like to be able to script it up a little bit because I'm probably going to have to do different combinations and don't want to be manually repeating it every time. Probably there'd be a great way of scripting all the connections, if I knew how!

Other stuff, though:

* Need to keep an eye on some notion of uncertainty / quality control. I'm not completely sure how I'll be doing that yet. Probably best just to get it working first with one approach then see.

## QGISsing up the intersections

In a Python script, so just gonna pick that up as I go along. Got it loading files - yay! Now just want to test the whole intersection process on a small selection. Which will involve having to select a rectangle. Wonder how I do that?

<http://gis.stackexchange.com/questions/86817/qgis-python-identify-and-highlight-features-programatically>

This has methods for selecting e.g. a certain area - but that provides a subset of features to then iterate over. I don't think that's what I want, though it may be...

<http://docs.qgis.org/testing/en/docs/pyqgis_developer_cookbook/vector.html#selecting-features>

Oh this is looking promising:

<https://docs.qgis.org/2.8/en/docs/user_manual/processing/console.html>

Nooope, that's only the processing items, nothing from the menus. Ah, this is the right question:

<http://stackoverflow.com/questions/27140520/how-to-access-qgis-menu-items-using-python>

Leading to this:

<http://gis.stackexchange.com/questions/57833/python-script-that-calls-the-voronoi-polygons-function>

Ah, no - nothing seems to be allowing me to use those (much faster, more flexible) menu items. This is puzzling... And is turning out to be a massive amount of faff. Might be better just to run them manually and deal with the fact I may need to do so again later. Annoying but more reason to make sure it's all documented.

..

OK, so back on this and really need to move forward with it. I think I should do what I was going to do before - experiment with one pair of polygons. So the plan:

* Run the intersect in QGIS because it'll give me the nice merged fields.
* Export for R so I can then better check on what the outputs actually look like, numbers-wise. E.g. what % are mostly fitting in, what are split, the cutoff for excluding what are obviously slivers etc.

So let's run! Actually, first thing I need to do: merge England and Wales. I'll stick those merges in their own folder. Vector/data management tools has a merge option - but the shapefiles to merge need to be in the same directory. So let's do that... in folders? Dunno, let's see.

<http://gis.stackexchange.com/questions/25061/how-to-merge-multiple-layers-to-one-layer-using-qgis>

* Merged England/Wales shapefiles, 2011 wards and 1981 EDs into own MapPolygons folder. Now I can intersect them and see what we have.

Right, we have the intersect! Now I need to look at it. Let's go all R on that mofo.

### Looking at the first intersect in R

File:

**Dropbox\SheffieldMethodsInstitute\Migration\MeenMigrationStitching\**

**analyseIntersectionFileFromQGIS.R**

**..**

Not far off 1 gig. Let's see how R copes. Well, it's slow to get going, certainly.

Right, so what do I want now?

* I need to get areas of all polygons stored in the same dataframe. Easy enough.
* Just looking at a little scribble of it...
* We need the % of the whole that each sub-area of EDs is. The area for the wards isn't necessary, only that the matching counts can get added to them.
  + I've got the ED area via gArea. So now I just need proportions of the sub-areas.

Actually, once it's past the initial stage, rgeos is pretty quick.

It's not very obvious what the cutoff point might be for dealing with slivers. Given they **are** slivers, they probably don't make a huge difference, though I can come back to them - think I'll need to do a select < area in QGIS to see.

For now, let's just check on the method for assigning %s of EDs to wards. Calculating %s first. Um... maybe lunch first!

## Back on census stitching

And starting with the geographies. Doing something simple before moving on to anything more complex. And concentrating on Scotland, not England. So I just want to get some geographies, then think about how they fit into each other. A little more experienced with the qgis/python stuff, though R may also be a route to go. There was some reason why I wasn't going to go that route?

Anyway - need reminding of the geographies. Walford was using census tracts, which aren't really...

..

So I'm back to reading through all the gumph in the evernote note. Also finding new stuff. It's all just superbly, hideously messy. So output areas explained in a little more detail. Note the point about certain variables - **which effing ones!** - not being available at OA level cos too small. Note the link to changes to OAs in 2011 at the bottom there ->

<http://www.neighbourhood.statistics.gov.uk/dissemination/Info.do?m=0&s=1457961571914&enc=1&page=aboutneighbourhood/geography/superoutputareas/soa-intro.htm&nsjs=true&nsck=false&nssvg=false&nswid=1680>

..

Quick walk. Things that occurred to me:

* Aggregating up to larger zones = win. Obv. For any 1971 voronoi EDs, we've got better odds of 100% matches.

Local authority probably too small in number, but we want to be able to test all of those fairly easily once the base data is there.

So, experimenting time. What do I already have / where did it come from / surely I've already noted that down, yes?

### Doing not staring!

This stuff is messy. I think the best way to move forward with it is likely to be just to work Witness-style and think/do. Which I'm doing at the moment.

File-wise, I now seem to be putting everything in the "data" folder. So I'm now testing in this QGIS project:

**C:\Data\MigrationWork\lookingAtScotsBoundaries.qgs**

..

What I really need: some re-useable code that is going to allow me to:

* Test intersection of various polygon types of different sizes
* Test interconnecting / inter-relating the different CoB categories - which is something it'd defo better to do in R, I suspect (or poss geopandas if I get to use the mac, though given that's not my main machine, perhaps not.)

Just want to check again on an example to understand intersect. I did do it above but now can't remember the exact details and would be a little surprised if R can't do what I need. But let's look again. Testing with some easy polygons:

* **2011 council areas and intermediate zones**

Actually, they're a bit too perfectly nested. Though you'd think that'd make it easier. Anyway: an error in R, I think due to nested polygons: "Geometry collections may not contain other geometry collections". And it took a loooong time to throw that as well.

Let's test QGIS again before anything else.

Also, I think I should backtrack onto a much smaller set of data (like I did before) to exactly understand what all the different approaches are doing. This is too slow. Thusly...

### Testing intersects

I'm just going to subset 2 IGs (intermediate geographies) and some of its 1981 EDs.

So the data:

**C:\Data\temp\MigrationIntersectTests\TwoIntermediateGeogs**

And the QGIS:

**C:\Users\SMI2\Dropbox\SheffieldMethodsInstitute\Migration\MeenMigrationStitching\QGIS\_scripts\Tests**

..

Really not being consistent with where I'm putting this stuff! Oh well. So we have:

* 1981 EDs: attribute table is label and name. Names are all NULL. Label is e.g. 6339AMO3
* IMs have loads more fields, include area and stuff. Err. How was I proposing to work out subsetted area, does anyone know?

OK, looking sensible. I should probably test if it works on the nested polygons that R was refusing to work with... Yeah, it works fine. Not that that's necessarily the problem. But probably is. Anyway.

I should say - having forgotten what I was aiming for / thinking the new areas were automatically assigned, it's straightforward(ish):

* Finding the area of the new list of polygons will then allow me to apply those to the attributes I want to then split across the others.

Order is important, or at least makes the results more convenient, depending on what's being aggregated to what. I'll work that out in a bit - I want to try again with coding it up in Python. Let's see.

**C:\Data\MigrationWork\Python\Tests\intersectTest.py**

..

Ah-haaaa. After some faff, we have succ... do we? Hang on... Yes we do. Now it's easy to check which way round things work. Thusly:

* The first shapefile to go into the intersection function gets the first ID column.
* So in this case we've got e.g. single-instance EDs that fit wholly into council areas, then others that I could find the area for that are assigned to one or more.

So we'd need the area of the **original** polygons as well in order to do the fractions. It's going to be more fiddly in Python to do this, I think. R:

EDs81@data$area <- gArea(EDs81, byid=T)

So it's a real shame about R not doing the intersect as I'd like.

Well. Hmm. I suppose I can just do all the area calcs in R. I'm just going to test that it is, actually, the sub-areas... ah no fuck it. There's no point, it wasn't working. Question: will QGIS work fast-ish on the lot? Let's test on larger amounts.

..

So: a whole intersect for Scotland (IZs into council areas) **took 7 minutes** in Python/QGIS.

..

Well. Brain is mush. Things to do next:

* I think, as before, it's worth really concentrating on a subset of polygons (7 minutes per test otherwise!) before moving up to the full thing. But some of that depends on what order I do things in.
* Somewhere down the line: centroids then voronois on **zones we know do perfectly nest** to get some sense of the range of possible misalignment. (And how much voronoi area might differ from parent zone area.)
  + Which will then let us get some range on possible outcomes depending on those numbers.

Tomorrow I think it might be worth thinking about the CoB links over the decades. What slots into what? Again, it's probably going to be worth doing two versions since 1971 data likely restricts all the other values too much.

### Doing some CoB work

Feeling a little more like I have a direction with this now. "Just starting" was the trick. Throwing self in. OK.

I'd already done a chunk on this but want to go over the whole thing again. I also think it's gonna be worth just starting the write-up somewhere.

So:

* SAS / Small Area Statistics: available for all decades. LBS (Local base statistics) only started in 1991 and, while more detailed, is larger areas. So we stick with SAS. (Then something else happens in 2011! But we'll get round to that.)

..

I did already put all this into a sheet. It's probably correct still. Let's look!

### Some little intersection tests

To put off looking at CoB matching for a little bit. I want to know if it's the nested polygons that cause R's intersection to fail. For which I will need a second set of polygons to test on that contain one.

Oh and also: **does** QGIS intersection actually do it correctly? Much easier to see on the subsets.

..

OK, so:

<http://gis.stackexchange.com/questions/151833/in-r-get-a-spatialpolygondataframe-from-gintersectionrgeos-with-sub-area-as>

The only option appears to be a LOT of arsing about converting to rasters. That does compute areas - but I can do that anyway by finding areas then working out fractions. I was going to ask a first stackoverflow question but think I'll get a little further on first. So the result of all that: stick to QGIS approach that's working, though slow, and... urgh... get on with CoB. Well that was a productive couple of hours!

### Back to CoB

This probably isn't that complicated. Ah, reminder: "super output areas" refers both to LSOA and MSOA. Obv.

<https://en.wikipedia.org/wiki/ONS_coding_system>

Which I'm looking at cos I want to know if the data is restricted at smallest scale again for 1991 onwards. I have a feeling that it's not, more recently.

Casweb for **Scotland only**:

* 1971/1981:
  + SAS only. Just written to UK data service, we'll see what they say.
* 1991:
  + LBS gives high detail but only at postcode sector or above. Not OA. Postcode sectors are close to, but a little bigger than, intermediate datazone IGs.
* 2001:
  + 'Census area statistics' option gives us the smaller range of categories.
  + 'Census area statistics univariate tables' gives us the larger amount. And we can get it for OAs. Bonza!
* 2011:

### 2011shmooshmoothandandeshmeven

Good luck reading that header. Scotland 2011: CoB categories are pitiful. I don't think there are any other sources. Just went to compare to England/Wales, which are amazing: 68 categories. (I've put an xlsx of those categories in -

C:\Data\Census\2011\CountryOfBirthDetailedAllGeographies

..

And these are the 2011 scots cats available:

* England
* Northern Ireland
* Scotland
* Wales
* Republic of Ireland
* Other EU: Member countries in March 2001 (1)
* Other EU: Accession countries April 2001 to March 2011
* Other countries

..

Not much good. Also: if I'm going to be asking them for commissioned data (which we might well be doing), is it worth seeing if we can get more detail on CoB for 2001 than is available via CASWEB for Scotland? How many CoB cats did we get for the univariate tables? ... 83. Loads, actually - so, more than we get in 2011. And different again.

So no real need to ask for more detail from 2001.

..

First-up: I should just check with them I'm right that we don't have very many categories before asking about more detail. I'm starting to suspect they're going to whinge about more detail...

OK, I've sent an email off. Oh Lord, it might take them ten days to respond!

What can I work on in the meantime? Some generalised code for harmonising the tables is the way forward - and a way to think about it. It may even be necessary to do some probabilistic stuff (I guess a little like microsimulation) but let's not worry about that right now!

We're still waiting on the reply about 71/81 - but I presume we're going to find out those are the only available categories. At any rate, this will be a simple-ish way to get started on thinking about linking. (Which is going to be easier for England, now, than for Scotland - until I hear back.)

..

There are probably a range of different ways to match for different decades too. It'd be good to have a clear way of doing that.

### Some quick notes on recoding/remapping variables

<http://www.cookbook-r.com/Manipulating_data/Recoding_data/>

And I've just been digging into dplyr, which I think I should definitely learn properly. Got a cheatsheet so I can get started with that.

### And then back to table stitching!

Which I suspect dplyr will come in very handy for. I'm also going to need category names for some of the CSV tables - can't seem to find a file for those anywhere.

..

This is an interesting source of census info:

<https://www.ons.gov.uk/census/2001censusandearlier/aquickguideto1991andearliercensuses>

Includes a link to a highly detailed census topic comparison across ALL censuses.

..

Also: this needs to work on the **actual** data. Now, if I want to be able to automate that at all and try different scales, I'm going to need to think about how that's coded up as well. Which will involve looking at a few examples.

The point being: I'm not applying summations abstractly, it has to be on specific data. I am staring blankly at it when I need to be getting on with writing down how to do it.

### Back to back to table stitching...

Yes on the last sentence: let's just piece them together. I had thoughts of automating the process with some kind of network code but I think doing by eye might be the way forward initially! Well, probably: with LBS level ones with sixty categories, that may be a little more challenging.

That said... I might just have a little think about the linkage idea. The point of it was to identify the largest contiguous span of matching categories.

First tricky categorisation:

* 1971: "N Ire etc" contains "other" areas of UK: Isle of Man, channel islands etc. Oh, but this doesn't matter, does it? We're only looking at cases of CoB **not** being UK. So those can actually all be summed regardless.
  + Only drawback: 1991 doesn't have "UK other", but it'll have to do.
  + Well. We can do England/Scot/Wales/RestOfUK. That's the only option that completely works.
  + Note again: 1971: the "rest" includes people who put "united kingdom". And we don't know how many that was.

Err.

* What does "old commonwealth" mean?
  + Guy, Will, and Ruth Levitas (Eds). Interpreting Official Statistics. Routledge, 2005 p.142
  + Old Commonwealth is Canada, New Zealand, Australia.
  + "New Commonwealth referred to those countries that gained independence in the 1960s."
  + Notes this was "tantamount to a vulgar distinction between white and coloured."
    - In fact according to this book, that was exactly the point:
    - Halsey, A. H. Trends in British Society since 1900: A Guide to the Changing Social Structure of Britain. Springer, 1972. P.453
    - "This division was adopted to give some indication of the numbers of Commonwealth citizens resident in the UK who might conceivably be expected to be white and those who might be coloured."
  + Ah: http://webarchive.nationalarchives.gov.uk/20160105160709/http://www.ons.gov.uk/ons/dcp171778\_352080.pdf
  + "The Old Commonwealth statistical grouping consists of four countries: Australia, Canada, New Zealand and South Africa." [33]
* There's also this worth noting:
  + "From the 1971 census it was possible to estimate the New Commonwealth population from the data on parents' birthplace which augmented the simple country of birth data."

Hum. I need a consistent way of marking uncertainties. Some categories are just not going to fit neatly together - we need some kind of solution to that, though I don't quite know what yet.

..

Hum!

<http://www.ons.gov.uk/ons/guide-method/user-guidance/longitudinal-study/data-quality/quality-of-census-to-census-linkage/quality-of-linkage-between-1981-and-1991-censuses--63-kb-pdf-.pdf>

And... there's another one for 91-01.

But they're both longitudinal study related, so let's ignore those!

..

### Commonwealth...

OK, so how to deal with this? ->

* 71/81/91: "old commonwealth"
* '01: "Africa", "Oceania", "South America", "North America".

So South Africa's in Africa. Oceania \*probably\* includes...

Well, given that most census categories appear to want to put darkies in their own category, it may or may not include Papua New Guinea. Let's try...

<http://webarchive.nationalarchives.gov.uk/20160105160709/http://www.ons.gov.uk/ons/guide-method/census/census-2001/data-and-products/data-and-product-catalogue/reports/definitions-volume/index.html>

So we've got Oz, NZ, "other oceania". In the main guide. Right then. So at least a small part of those are Papua New Guinea - but we're probably not getting a large number from there, huh? Guess I can compare to the larger LBS breakdown to see.

Actually, for 2001, univariate tables, there's more info. Sooo...

Actually, yes: I have more cats, which can help with the americas too. Need to write those down also.

..

### Rethinking strategy

New solution:

* Get high detail data for 91-11. (~~Which will mean working with England again for now if I want to include '11~~Update: newp, fine for Scotland!)
* I can then backcast those categories better to the remaining 71/81 data. Things like "far east" etc.

~~We'll then have to wait on Scotland data. Could be a little while. Or I could get Scotland done for now up to '01 seeing as I said I'd do Scotland, and we don't know what categories are coming next. Err. Time for a walk? Time for a walk. // Walk done~~. Now there's this. So.

* 1991 geog that allows LBS: Scotland = postcode sectors, England = wards. They're probably about the same average area (though av pop would be more relevant obv).
  + Ah no: much larger for postcode areas, but some of those are covering very sparsely populated bits of Scotland.

OK, so. What do I need to be doing here? Well: '91, '01 categories, back-casting to 71/81 in Scotland.

..

Have typed out all of 1991, I've just realised it's obviously much quicker to copy/paste. So I've copied 2001 to a doc on the desktop. Format not complete. At some point I need to go back and get the codes as well. OK then.

### Having got detailed 91/01

Which annoyingly means only large-scale for '91, so affecting all the other geographies.

### Scotland!

In the Output Area bulk download: Table QS203SC - Country of birth. Has all the categories. Doh. Oh well, at least I checked!

I should probably attempt stitching the buggers really. Let's add the new Scotland cats to the table. I should probably check if they're different from England's.

## Restate assumptions

Need to recount where we're at So I can tell Gwilym where we're at. Two specific things:

* Process of geography aggregation. I am getting somewhere with that, though perhaps not quite as far as I think I am! All I've managed to do thus far is identify the correct way to python up the intersection. I still have to work out areas etc.
  + And I'm leaving out anything more fiddly like dasymetric mapping, at least at the moment.
* Recoding CoB categories: restrictions across censuses.
  + Smallest areas: EDs or output areas, depending on year.
  + I'd been atttempting to aggregate up to intermediate zone in Scotland / MSOA in England.

On 1991: difference in size between postcode sectors / OAs:

* ~5000 average population for Scots 1991 postcode sectors (1000 zones)
* ~130 av pop for OAs. (~40K zones)
  + Probably skewed by urban sizes, but...

Ah, little aside: **population bases are different for each census.** Which I knew from Dorling, right? But you can see it in 1991 "population bases". Well, for now, let's look at "all ages" to get some rough numbers.

..

Update: population bases might not matter -> Meen's model is about distributing the **total migrant group share** for the country to different zones. That said: at some point if we're thinking about the underlying spatial forces, population might come back into it again.

Oh and, yeah - actually, it's there in the labour/housing market vectors: e.g. population vs housing density.

## Stitchy stitchy

Back on this after a hoofing amount of work on housing data. A deadline for producing actual output in the Summer - 23rd June presentation. Certainly focusing the mind! Now remind me, where did I get to?

Well, I've read through it all and it's all still seeming as murky as ever! Reckon it's just gonna have to be another case of "start". Pick a point, do it.

I do at least now know the basic restrictions across the five decades in some detail.

* Think it'd be best to shift the Rstudio project over to the C drive - in the Census folder.
* There are bits of code in the Dropbox SMI/migration folder that I should probably go look at.

For now - need to actually try and do this. First thing's going to be getting the actual country names from the column codes. My googlings failed to find anything for that before.

### Quick WICID point

Was mulling again its use to check on CoB changes being due to overall increases or movement: CoB categories are too crude. And also possibly only available for the last two censuses.

### And then back to stitching

And I'm going to start trying to stitch. I could probably do with some systematic way of doing this, but... well, I suppose some straightforward name-based function will probably do the job.

All I need is some way I can systematically re-code.

I should probably think logically about the stages here too - moving on to how I should use the info coming out of the Python geography data. Though I'm pretty sure it's logically entirely separate, isn't it? Because ->

* Category-combining is summing columns.
* Geography-combining will be summing/splitting rows.

I'll be needing some sensible re-coding first. And do I have everything I need yet?

### Martin/Dorling

Ha: ``While each of these inconsistencies arises from quite understandable circumstances when considered individually, the combined result is to present an almost impenetrable barrier to the social scientist wishing to examine small-area social change using standard census outputs.''

(Martin / Dorling / Linking censuses through time, 2002)

There was a "linking censuses through time" project - and at some point a website, though that no longer seems to exist.

* Note this too: 1991 census apparently missed ~100 people per ward. Not clear how they know that! But...
* P.86: is saying that **correcting for 1991 mis-counts** was the largest amount of work. Errm. OK!
* It revised Scots counts up by 2%. So I suppose it depends if that kind of level becomes important, or how to add uncertainty.

### And back to stitching again!

I was trying to find variable names. Think I'll just have to do it manually...

Which I think I'll do with an eensy bit of logical code and staring. There's... oh man, I was looking at beautiful soup. Stupid idea.

Things to note:

* Where I can, just get univariate/non-female/male split data.

..

### Column processing

Summing male female etc.

All going OK. One thing to note:

* Removing sum columns: 1991 SAS for 'United Kingdom' seems to have Isle of man and others hidden inside its sum - the nested items don't perfectly sum to it. Soooo need to go through and check if that happens anywhere else. I don't think so, but here I can turn it into it into its own column.
  + Actually, they've just as often got more as less. So I think just ignore that column as we were doing before!
  + I've save that as UK\_total\_diffToSeparateColumns.xlsx if we need to come back to it.

..

Notes!

* 1991 LBS postcode sectors: the data has a **lot** of completely blank rows for certain postcode sectors. Why? Dunno. Will have to check when mapping.
* Well: just downloaded boundaries **with** the 1991 LBS data: they're different postcode sector boundaries, not the 1991 ones. So there's that. Shouldn't be a problem - and number/name of zones is always a good guide to whether it's the correct match anyway.

OK, that's all five into some sort of tidyness. Now for.... **lliinnkking**.

### Stitchy stitchy

My section names are not amazingly descriptive, are they? Oh well. So: we have all the data in something like a neat order. Now I can choose some stitching. There are two basic options, with a possible third that I'm mulling but haven't worked out:

* Across all five: minimal categories, restricted by historic.
* 91-11: a lot more categories but '91s LBS only to postcode sector level (1000 zones in scotland, or less given how many of those are blank.).

..

I should also, possibly, remind myself what happened with the area re-assignment so I can think how to apply that. Let's do that right now!

### Trial run of data re-assignment linked to zone re-assignment

Which is in C/data/census:

**intersectionTesting.R**

Hurm. Wot I fink: might be worth trying to do the basic geog thing, just once, so's I feel I know it's doable. So for instance:

* 71 EDs to 2011 IGs? So I need:
  + The overlap shapefile from GQIS/Python (71 into 2011 I think is the correct direction)
  + Load into R to get actual areas of those.
  + Attempt re-assignment code!

### Having a go at the geog stuff

First hitch:

* There generally seem to be more geographical zones (for 71 and 81 at least) than there are zones in the CoB data.
  + This is the right way round to be able to work with though: Any re-assigned people will still work. But it would be good to see what zones are on the map that are **not** in the CoB data.
    - Which you'd think would be easy to select but, for some reason, it's hurting my head. Or, it doesn't work with spatial data frames. Let's check again...

<http://stackoverflow.com/questions/13443372/simple-way-to-subset-spatialpolygonsdataframe-i-e-delete-polygons-by-attribut>

I think the key comment there is likely to be "R doesn't automatically drop factor levels". So.

..

I'm being very silly. Some logical code is doing exactly what I want it to. OWOWOW! HEAD!

### Spatialpolygons shizzle

This:

* Zone IDs in multiple rows in '71 polygons (and I suspect '81 too)
* They're identical polygons - but multipart. Or should be multipart. So mostly round Scots coasts.
* I can do this to them:
  + <https://philmikejones.wordpress.com/2015/09/03/dissolve-polygons-in-r/>
  + Then just turn back into DF before saving as shapefile (cos only need ID, there's no other relevant data)
* But I just want to check this isn't a problem for later polygons. Do e.g. 2001 output areas do the same thing, or do they have multipart?

..

So there's one that's got the correct number (after removing some fiddly rows!):

* 2011 OAs match in both data and zones. Yay! And will probably... yup, nest perfectly into intermediate zones.
* Though actually... using those means each fits perfectly - it won't let me test splits. Hurm.

OK, I'm trying it instead going back to 1971 EDs. I've completed the dissolve/re-assemble (which would be nice to write up! I should get practiced at that, it might be useful. If it wasn't quite such a pain. Maybe I could try later.)

Tick! All good, checking in QGIS: the previous duplicate IDs of e.g. polygons either side of an estuary that should be one multipart - are now one multipart.

So let's use this and see what the zone/data ID match level is like...

### 71 matching (not quite)

This:

* There are 111 more rows in the CoB data than geog zones. I've double-checked, they're pretty much definitely not in the zone data.
* So no real option but to drop those. Out of ~15K zones, not a huge issue, but a puzzle. One to note and chase later. Hmm... I think I need a 'snags' table!

Just re-running the intersect using the dissolved '71 data.

Right, so - remind me what I've got and what I'm aiming for again?

* 71 ED-to-2011 IZs:
  + The intersect shapefile gives me ->
  + Possibly multiple '71 ED IDs: each one gives an area assigned to the larger IZ.
* I need to turn those areas into fractions with the area from the parent '71 ED.

..

OK, getting there with some fiddling. I have my fractions - let's apply them... after lunch!

### Applying fractions

So that's:

* Splitting up the various CoB cells into the larger geographies
* Then summing them back for new totals

Err. I think I just need to merge the fraction (with the OA codes) into the CoB data. Then sum by OA code. I think that's it. So that's...

..

Working! I think. Exported to QGIS for a look. Had to do this:

* Number of people / area of polygon.

Somewhat obviously (num people per m2). But QGIS makes that really easy.

..

## Where is this shizzle at?

Well.

* I've tested all the basic components and everything works.
* But I've only tested for 71 into 11. I need to think through doing this for everything.
* And that includes applying the same to the other census variables - and just actually getting hold of those is a large task in itself.
  + Relatedly, important to think through: how all this relates to census population. There's stuff to do with census pop bases I suspect I don't understand.

So each of those census vars needs to be put through the same process of splitting and combining. I think I should be doing som functional coding for this, or it's going to get messy. But how to boil it down?

One quick thing I want to check on before continuing:

### Keeping polygon intersection area code in R

So the whole thing can be kept internal. Just thinking ahead to making the code available. There's a raster way, I think.

<http://gis.stackexchange.com/questions/140504/extracting-intersection-areas-in-r>

..

I feel like I've been round this before but I want one more try from scratch. Ah: so far, I'm remembering that R didn't like something about some of the shapefiles. QGIS/python doesn't complain. May have found a fix but let's see. Oh ->

Data/Census/intersectPolygons.R

Oh Christ, it's taking forever to run! Think I might try that a little later on. I need to be doing other things!

### So putting it all together

What have I done, and what do I need to do?

* Geographies: intersected 71 EDS into 11 IZs.
* 11 IZs is the "shared geography" I've currently chosen - all other geogs will go into that (with 11 OAs nesting nicely).
* Areas of 71 intersects, as fractions of whole area (find after intersect cos of edge issues etc).
* Use fractions to multiply through all census variable columns then re-sum into larger geography groups.

Let's try and generalise that mentally before thinking about applying. The only really fiddly thing I want to get in here:

* Applying the same to **an arbitrary number of columns** in a given census dataset. So I suppose just passing the dataset in, the ID column, the positions of the columns to multiply through.

Sub.Building/Building.Name/Property.number

### Going through the code

Things wot I did ->

* Prior to doing the main thing: make sure all small-area polygons are one feature per row, i.e. any zone ID that's split across multiple rows, turn into one row with a multipart polygon.

That's a zone shape prep point. Then the main jobs:

* Sanity check on zone number:
  + Is there data for every small-area zone? If so, continue. [If not, will mull!]
* Find area fractions via areas for the small-zone slices over totals for the whole small zone:
  + But use the area total from the intersect shapefile (group by / sum). It will already have been cut down by the larger geographies edges and so will be more suitable. (Probably. Was in the case of 71->11)
* Merge the resulting fractions into the census data: so each small zone now has multiple fractions attached to it.
  + Multiply the appropriate columns through by fractions.
  + Sum back up to the larger zone.
* Turn back into SpatialPolygonsDataFrame, save as shapefile.

..

So, turning that into a function?

* Out: SpatialPolygonsDataFrame of new values assigned to a larger zone set.
* In:
  + Intersection SPDF, pre-done in Python/QGIS (don't need individual ones)
  + Census CSV data with matching small-zone row IDs.
  + Probably details of ID col positions and census data positions

Oh: I need to do my checks that the small zones are one row per feature elsewhere: they're obviously multiple entries in the intersect file.

Ah ha!

<http://stackoverflow.com/questions/21815060/dplyr-how-to-use-group-by-inside-a-function>

That's what e.g. group\_by\_ is for. Gotcha.

..

OK! Well. The function turned out annoyingly different. Close but not exactly the same. Why not? I can guess some possible reasons...

Oh, that's confusing. Row name order is different. I can see this being fiddly to track down! Brain is cooked in 27 degree heat, might have to do something else.

..

Note to self: people per square KM is a **stupid** measure. It should be e.g. "English as % of pop". Really really obviously. How embarrassing! Well, could have been worse.

Also note: natural break choropleth in QGIS never \*quite\* produces the same numbers for the same data. Very odd behaviour - I mean, actually loading the \*same\* layer in twice and doing it, you get a slightly different result.

..

At any rate: first pass of the function, it's working. Next-up, on the morrow:

* Set up QGIS/Python properly to supply the intersect sets (folder-full of smalls to go big.)
* Get some of the other data (I need pop data!) to test on. The same process has to be run for all those datasets.
* And the big thing: deciding on CoB re-categorisation and applying \*that\*. (Which can be kept quite separate: summing columns doesn’t interfere with summing rows!)
  + But this should probably include two options: 91-11 and 71-11, with diff cats.

I also want to be thinking about uncertainty a little. There might be a way to think about split areas and where people might end up in each - and can then do a range of answers.

Actually, that really might work. Totally contained ones, you know the result (though also, need to check on voronoi result diffs!) Partial ones will be some split between various parts. There's an approach there.

OK, that might do me for the day.

### Back again

Python: later I should probably code something that will take in all shapefiles from a folder and intersect them with another (so we can try with various geographies, test differences etc.) For now I reckon I'll just run those four through manually.

First, need to check the other small-scale zones have a single zone per ID row. Another function?

Done! Just tested on ed81s, looked at in QGIS: works as advertised.

Now going to run back through the QGIS/Python intersect code. Take a note this time of how long it takes to run - I still want to compare to R-native options, though may have to run over lunch...

* **81EDs -> 11IZ intersect in QGIS/python: 3 minutes**

..

OK, all working still. Functions all set up then. Still messy because various sanity tests required. But '81 EDs looking good. Now I have to stop with all that for a bit and move on to two other things:

* Decision on aggregating CoB columns. (Two timespans, probably.)
* Getting the other data.

The latter's probably a priority: I need population data to be able to do e.g. "English as % of zone pop" calculations. And I need to figure out a neat(ish) way of doing that once population's been through the intersect process.

Along the way, I could also do with calculating what number of small zones were 100% in the right large zone, which got split. Those numbers could do with graphing up. But for now:

## Get rest of census data / apply population data to get proportions

A thought. Do you get the same result if you do this?

* Find proportion (e.g. English % of total pop) in each small zone **then** aggregate via intersect.
* Aggregate CoB and population separately **then** find proportion

I think it should be the same: the same proportions of each will end up... yeah, it should be the same! But might be worth checking. Hang on, I can do this on a piece of paper.

Ah no: you can't just use proportions. Thx paper - obvious point but: e.g. 3 small zones, 10%, 25%, 15% English - you can't aggregate proportions to a larger zone because you're not adding like to like. 10% of 100 plus 25% of 250...? -> you need the underlying pop number.

Oh, but you could find out the... noooo, we already **have** the exact number. Oh man, sometimes I could...!

OK, so that's easy: sum everything to the larger zones **then** proportions!

Note the problem here:

* **Population is likely not distributed the same as sub-categories.**

But we use as small zones as we can and hope for the best. It's only a problem for split zones. I'll do some counting on that.

..

### Mulling CoB

Another thought knocking about:

* How are we to deal with the basic uncertainty surrounding CoB stats? Given we're trying to examine movement of people. From CoB alone:
  + We don't know if a head counted is someone who **hasn't moved** since the last census; who **arrived from abroad** since then (so a net addition overall); who **migrated internally**.
  + We only have the stock. Geoff points this out too, but I think it's especially problematic for the first category - people who didn't move.

Something to come back to! I think for specific censuses, there are likely ways to deduce what these numbers are by looking at the spread of CoB versus "where I was a year ago" (for those censuses that I can get that data for). It might stay reasonably consistent, which might help break the CoB figure down into some spread.

..

Then thinking about what the Geoff model's actually got in it:

* % CoB out of 100% total for area, so in this case presumably Scotland.
* Looking at it, I'm all over a little confused about some of the shares/percentages etc. Population size of the zones in question **is** in there, but in a roundabout way (via local pop / number of dwellings).
* So keeping raw CoB numbers = important. As well as making them a % of total for whole of Scotland for each group but that's doable per table.
* Note how that does away with any endogenous change in the **total** CoB number. Perhaps that's good / what we're after, but it's confusing me a little right now.

BUT! For proper maps of raw CoB -> it needs to be e.g. "English CoB as proportion of zone pop".

And totals should add up shouldn't they huh?

### 1971 population

Right, so... population data then? OK! Should already have that for 2011. I know it's already attached to some of the shapefiles but I think I'd like to have the raw numbers too.

1971. Got the numbers. Don't understand them. Let's see if this helps. Note the googling term is "census definitions".

<https://census.ukdataservice.ac.uk/media/51156/1971_defs.pdf>

..

OK, so we're going to have different methods of counting people. This one, for "enumerated people", is eeeveryone who was there, resident or no. Campers, Vagrants, their dogs, everyone.

Actually, I think the first column there must be the main figure. But worth coming back to check at some point. Let's just assume that's right for now! So:

* **1971: use first column in pop table.**

### 1981 population

Ah OK: look at the "population base" table. It shows the difference:

* 1971: it's everyone and their dog who's actually there. If they live there and are away, they're enumerated wherever they are.
* 1981: it's everyone who lives there **even if they're absent** (but not all visitors in the area, which makes sense that's be exclusive).

So we can at least tell from this table what the scale of that difference is. I'll have a think about what to do about it shortly!

Let's see what happens in...

### 1991 population

Which we know from Dorling is a bit problematic. Can't remember why. Let's see what's there. I wonder if the 1991 SAS/LBS pop tables are different too?

Well. We've got 71/81 bases again! Worst comes to the worst, we can look at distributions of difference to model uncertainties.

### 2001 population

Note: in key statistics / usual resident population -> we've got a bunch of footnotes on issues with enumeration. Useful but a little tricky to find out. (None of the other population offerings have such detailed notes.) Key parts:

* "Care must be taken when interpreting the intercensal population change, as there have been changes in definition between 1991 and 2001, and there is a small element of under-enumeration in the 1991 counts, the 2001 counts have been adjusted to account for under-enumeration."

Probably need to go read the Dorling stuff on this huh? For now... dooownloooad. What to pick? I should probably compare the various options. So:

* SAS table CS001: "age by sex and residence type" has an "all people" option. Let's see. Has the advantage of being one straightforward number!
* Univariate statistics / population density:
  + Yup, those two last are precisely the same number. Good!
* Key statistics / usual resident population: which contains 1991 population too, adjusted - no, not adjusted, I don't think, but counted in the same way. (2001, like we just said, **was** adjusted for under-enumeration.)
  + Oh: the 1991 field's redundant. Same figures all round! And same as the other two sources. OK then.

### 2011...

We already have via the bulk download. But what's the table name? Possibles:

* DC1104SC: err. No such table in bulk download.

Ah:

* KS101SC: usual resident population
* KS102SC: Age structure
* KS401SC: Dwellings, household spaces and accom type (for housing density)
* QS101 and 102, 103: residence type, pop density, age by single year
* QS112SC/113SC: household composition by people/households
* QS417SC: household spaces
* QS418SC: dwellings

Note also:

* LC/DC2802SC: CoB by age of arrival in UK. (Err. But how old are they now?)
  + No we don't: it's **only** age of arrival, like wot it sez.

### Bigjobzz: deciding on CoB category linking

Better get on and do this one! Just needs staring at and figuring out. With the caveat of doing it twice and running 1991 geographies a little differently.

So a bunch of them should stay more or less the same. Some are going to be a little odd. Let's see.

..

Note, for thinking about 91 differences:

* Postcode sector is ~65km^2, OA is ~2km^2. (miscShapeFileExamininz.R)

..

Right-ee-ho. We need to get these CoB categories done. Again, dealing with it functionally might be better: I'd be able to work with it a little more systematically. Hopefully. Well...

* I've got all the names in an excel sheet: in sheet 2, let's remove the ones we've dealt with as we go through. So to start with, a number are common across all...

Thusly:

* England/Scotland/Wales/Republic of Ireland removed.

### "Rest of UK"

Some options here! N Ireland not cleanly categorised. What differences across the five/six? (And what were the notes about UK other and all that? Need to open that sheet again! Diff location - dropbox/sheffieldmethodsinstitute/census...)

* 1991: the larger category options give us a way of checking the differences in Norther Ireland / Rest of UK / Channel Islands etc to get some sense of the magnitude of the difference.
  + 1991 SAS otherwise: hard to know where those numbers end up.
* If we use 91 LBS, it would be possible to add channel islands / isle of man over the whole lot into a consistent category. As it is, it's difficult to know where they've gone.

OK, going to make this (probably incorrect) assumption for now, then use SAS/LBS difference later to check:

* 1991SAS N Ireland **includes 'rest of UK**' .
* That then allows for the first re-codings...

So:

* 71/81/91: N Ire / rest of UK is now a single category: **leave as is**.
* 2001: four categories to one "rest of UK":

Northern.Ireland

UK.part.not.specified

Ireland.part.not.specified

Channel.Islands.and.Isle.of.Man

* 2011: 3 to 1:

Europe..United.Kingdom..Northern.Ireland

Europe..United.Kingdom..UK.part.not.specified

Europe..Channel.Islands.and.Isle.of.Man

### Old commonwealth

"Australia, Canada, New Zealand and South Africa." (see above)

* 71:91SAS - 'old commonweath' present. Though: South Africa is 'new commonwealth' in 91LBS. Oz/Canada/NZ is old.
* 2001: loads more categories, of course, so:
  + Oz/Canada/New Zealand to "old commonwealth"
* 2011: Same.

Wow, this delete method is really working!

### Africa

* 71: just Africa
* 81: 2 cats combine - eastern / other
* 91SAS: same, 2 to 1.
* 91LBS: see cats in CASWEB, all subsets of SAS so easy to re-code if needs be.
  + But broken into two groups: non "new commonwealth" ones further down.
* 01: see CASWEB again. Though note:
  + North Africa is its own category. Don’t miss that. (Update: doesn’t matter right now, re-coding to 'Africa!')
* 11: all nicely labelled "Africa" to start with.

p.s. in 1991, this is "other new commonwealth" (so no major African countries):

* "Ascension Island, British Antarctic Territory, British Indian Ocean Territory, Brunei, Christmas Island, Cocos (Keeling) Islands, Cook Islands, Diego Garcia Islands, Falkland Islands, Gough Island, Heard Island, Kiribati, MacDonald Islands, Namibia, Nauru, Niue Island, Norfolk Island, Papua New Guinea, Pitcairn Island, Raratonga, Solomon Islands (British), South Georgia, South Sandwich Islands, St Helena, Tokelau Islands, Tonga, Tristan Da Cunha, Tuvalu, Vanuatu, Western Somoa"

(from definitions 1991)

### India/Pakistan

1:1. Are their own countries in all of them. Nice. So let's remove those.

### Europe

* 71: 'Other Europe' is a single category. 81/91: Europe/European Community all have to go into that big pot.

The only tricky one here:

* 91LBS: Turkey sits in its own category. (As does USSR)
* In 2001, it's in Eastern Europe. (And Europe/Non-EU in 2011).

So where does it end up in 91 SAS? It must be **rest of world** if it's not in Europe, eh? I can check if the sums work at some point. (That'll be a laugh!) So:

* **Exclude Turkey from other Europe cats** to make consistent. Rest of world! Does that work?

OK! Europes all bagged together! (Except Turkey, who we're kicking out into purgatory for the sake of consistency.)

..

Note: '81 Mediterranean is Cyprus/Gibraltar/"Malta (including Gozo)" (from the 1981 census definitions)

### Far East/China

<https://en.wikipedia.org/wiki/Far_East>

East Asia.

This one's a bugger. I can do something but only by working with 19 LBS data. Otherwise, there's no way to create a consistent "Far East / East Asia" category across all decades. Tea!

So:

* 71/81: Far East (=East Asia)
* 91 SAS: CHINA! Damn them.

Which means:

* China is a subset of 'Far East' but we don't know how small. The only common category: "rest of world". Not ideal.
  + So that's losing detail on the entire of East Asia. And possibly South East Asia...
  + Oh and just to point out, there's a principle here: you must be able to find a common bucket to stick them in. If they're **already** in thesame bucket, all the better.
* I can use 91-LBS as a way round this, but it's not ideal: the geographies are large and it introduces a lot of uncertainties.

Yup, rest of world binning for this one! For all five at least.

### Remainder of New Commonwealth (only good for '81 on, it turns out)

From '81 on: could do this. It's Cyprus/Malta/Gibraltar and a couple of others. But nothing in '71. (All has to go in 'other').

### "Other"

So in sheet 3 of

C:\Data\Census\VariableCoding

1971to2011\_CountryOfBirthCategories\_Scotland\_noRegionHeadings.xlsx

->

**All of these** now have to be "other". Now I have to think about the other versions of this we can try.

### The tricky ones...

Quick summary of the above:

* If I do different time-periods, Europe, Africa and the East all have different options.
* There are even different options if leaving out '71: **Caribbean and Bangladesh** are in play from '81 onwards. As well as 'New Commonwealth remainder' which seem to be a common group of Cyprus/Malta/Gibraltar and a couple of others.
* Re-coding some of the cats in 91-LBS provides more options, but requires some extra geographicalness. Well, it doesn't - it's precisely the same process, isn't it? There's just more uncertainty about its accuracy. (Yet to be assessed.)

## Thinking through how to carry that out

I want to do a little planning here. It would be possible to do the above 71-11 re-coding and press on manually, but it's going to get unwieldy and hard to follow when there are several re-coding versions. I want all that code to be making sense.

And just looking at the geographies:

* 1991 Postcode sectors are a very poor match to '11 IZs (entirely unsurprisingly!) and are often larger. Err... ~61km^2 mean area. (Postcode sector is ~65km^2, OA is ~2km^2 - so yes: postcode sectors are larger.)
* Obvious solution: **aggregate geographies to 1991 postcode sectors.** We can then use all 1991 LBS data. Disadvantage: not at all designed to match... that quote:
  + "Where possible, they have been made to respect physical boundaries and natural communities. They have a regular shape and, as far as possible, contain households with similar social characteristics." <http://www.gov.scot/Publications/2005/02/20697/52626>
  + Though who knows exactly what that means?

Well that makes a lot of sense, doesn't it? Consistent zone in the middle of the time range. Trade-off: no perfect fit for 2011 categories (and to an extent 2001 as well).

This leads us into plenty of options for running for all and seeing what error we get. But **a big problem for 1991 data just goes away**. We like it when that happens.

So! That leads to three sensible options to try. No it doesn't. There are more. Let's see:

### Base geography: 1991 postcode sectors

Three aggregation options:

* All five censuses: I can now do a little bit more with "far east". Everything else still needs "othering" (ahaha).
* '81 onwards: far east, Caribbean/Bangladesh, New-commonwealth remainder can be played with.
* **Much** more detail for '91 onwards. I'll try that in a minute.

Postcode sectors look a lot larger in Glagow city. They're likely to have a much large population range, I suspect, too: OAs to datazones - aim to be homogenous, don't they?

Well: I think now I've realised that, it's probably a good idea to press on with this option for now. There are interesting things to be done to test with 91 data since it has both. But for now: creating three different sets of categorisations seems like enough to be getting on with!

### Reminder of the whole process

So I can think where to put things as I'm going along. Soooo:

* Blue Peter data: "here are the zone intersects I made earlier" in QGIS/Python. Which will now be everything into 1991 postcode sector.
* Get census data ready for mapping between zones.
  + CoB: finding common CoB categories across decades: sum those columns!
  + For which I'm going to do three different options, so should use a folder for each.
  + This is going to require making sure there are common names for those groups. I could do with a way of making sure that happens! (Or I could just hideously headbutt it).
* Chuck all those through the re-assign function mangle. Voila: stitched.

Well - I'd better get on with the re-assign. Um.

Actually, yes: I could do with a function for re-assigning name by index, to make sure they're all precisely the same. It's a little fiddly to know how to do; I could do with more R knowledge at this point. I want it to be able to work with 3, 4 or 5 sets of data.

So a function doing this:

* Object that gives:
  + X dataframes
  + Common name for all columns that will be re-coded (or left the same)
  + Indices (or names to be sure?) of columns from each dataframe to be combined into that common name.

So it could be a list. OK... got enough to test some code with.

### Function for re-assignment

OK, it's turning out to be a little bit fiddly. Probably best to try and write about what I'm exactly trying to do. Function is doing this:

* Takes in list of dataframes to process
* Then lists with a new variable name... hum, might be worth making those variables first huh? Or not. Err.

The basic principle here isn't tricky: just selected the appropriate columns, then apply a sum, then rename, then stick on the new one. Repeat. Just gotta work out the cycling order...

Might be wise to return to this in the morning. This is going to take all week, isn't it? Harrumph!

Just running tests... well, I guess R will open in the right place. Just the dataframe test in the function.

..

Huh: note here "rollapply" in quantmod. I've wanted to compare previous values quite often...

..

Looking for a functional way to make a dataframe. Many options...

<http://www.r-bloggers.com/using-apply-sapply-lapply-in-r/>

<http://www.stat.berkeley.edu/~s133/Docall.html>

Let's just run through the logic of what we're doing again:

* I'm passing a list of dataframes
* Then a bunch of lists, each of which codes a new variable name and groups of column indices to sum into that new name
  + So each of these variable names: it needs to act on each of the dataframes in turn.

Which is the part hurting my head: working out what order to iterate in. It \*probably\* makes sense to make each dataframe in turn - which means

* iterating over dataframes first,
* **then** each of the var lists (with the list index+1 corresponding to the dataframe columns to sum)
* This way I can make all the columns for a single dataframe then add it to a list.

Sooo. I'm not quite sure if that means I'm making dataframes to start with. I think I can just make columns then maybe rename afterwards.

I'm trying to use lapply: so a list and then a function applied to each element of that list? From which I want a bunch of columns, one for each re-coded country? So...

* The list to apply the function to: this can presumably be supplied by a function if needs be. What needs to be in it?
  + The cols to sum for each country. So just that value, if I can pass that to a function to calculate. (Can rename after).

...

Getting there: Just got to do some adjustment to column indices. I forgot about ID columns etc. Thusly...

Hmm. Seems to be working. Need to tidy a little.

..

Error! Took a while to track down. The 2011 data - it's got some non-numbers in that buggered things. Let's see what... ah, commas. Right. Well... now, it does in fact seem to be working. It is! I've checked! We have sums! Exclamation mark! Blimey, that was some challenging coding. Felt like I learned a lot though. Errm.

## Using the recoding code

p.s. On Friday let's do this:

<https://hilaryparker.com/2014/04/29/writing-an-r-package-from-scratch/>

via

<http://r-pkgs.had.co.nz/intro.html>

Then get it up on github.

..

This'll need githubbing soon. But now we're back on to actually doing the re-coding. Which we worked out for a first pass, didn't we? Oh, except - a couple more options from using LBS. Which were...?

* All five censuses: I can now do a little bit more with "far east". Everything else still needs "othering".
* '81 onwards: far east, Caribbean/Bangladesh, New-commonwealth remainder can be played with.
* **Much** more detail for '91 onwards (and bigger job)

So let's get everything working across all five for the easiest. The only change was to Far East... so let's do the easy ones first (ones I've already been over!) then re-do that.

..

Some updates:

* USSR is in its own category mostly. Others have it as Europe. I'm gonna say it isn't because we're coding to 1971.

### Back to Far East

This one I'm re-naming East Asia! And we have to decide what's in it. Oh, except it may not just be East Asia. Let's check 'definitions' again.

Wow: the 1971 census was **hugely** racist/colonialist! It turns out 'far east' is a 'sub-division of the new commonwealth' that doesn't include China (p.247). We've got:

* Hong Kong, Malaysia, Singapore, Brunei "and all Commonwealth Islands in the Pacific not elsewhere classified."

So China is in "rest of world". Huh. And the choice of census classification was entirely about cataloguing Empire flows. Huh. Is it the same for '81? (I should save these and stop googling!)

That's annoying: the 81 cats in the actual definitions are actually way more detailed than is available. Someone must have re-coded to '71 (mostly) so we'll have to assume it's the same countries.

So what does all that mean?

* Only a small number of countries get to go into "Far East".
* A lot of others will **still** have to go into "other".
* And I'm going to keep "Far East" to make clear what's going on, maybe put in speech-marks!

Hmm. This is actually rather more fiddly than it appears...

### New commonwealth

We have two specific problems here:

1. I've dumped countries from Africa into "Africa". It's wrong, I'm pretty sure - "Africa" in '71 is "new commonwealth Africa". Other countries are separate. They **were** present in the original categorisations... oh, in 1981 ones, not 1971.
2. 1981's original cats have been changed, probably to be closer to 1971, but follow the same basic principle of "empire vs everything else".

1981 gives us a list of (at the time) **non-new-commonwealth African countries**. These I will need to dump into "other". They are (from the definitions; we get none of these in the data):

* Algeria, Egypt, Libya, Morocco, Tunisia; South Africa;
  + And "other Africa (foreign)", which is: Angola, Benin, Burundi, Cameroon, Cape Verde, Central African Republic, "Ceuta and Melilla", Chad, Comoros, Congo, Djibouti, Equatorial Guinea, Ethiopia, Gabon, Guinea-Bissau, Ivory Coast, Liberia, Madagascar, Mali, Mauritania, Mayotte, Mozambique, Niger, Reunion, Rwanda, Sao Tome and Principe, Senegal, Somalia, Namibia, Sudan, Togo, Upper Volta [now Burkino Faso], Zaire.

<https://en.wikipedia.org/wiki/Ceuta_and_Melilla>

Canary Islands; Spanish Morocco.

<https://en.wikipedia.org/wiki/Comoros>

Huh! Didn't know that was there.

<https://en.wikipedia.org/wiki/R%C3%A9union>

<https://en.wikipedia.org/wiki/S%C3%A3o_Tom%C3%A9_and_Pr%C3%ADncipe>

So just to be clear:

* **"Africa" 71 and 81 do not include any of those.**

### "OTHER" New commonwealth

Which is going to vary across decades, but let's try and get this one out of the way. So: we need to throw Sri Lanka into this pot in 71 cos it can't be linked in 81.

This is where we then need to separate out

* "America new commonwealth" and "Far East new commonwealth".
  + America new commonwealth need to go into "other new commonwealth": again, nothing to match on in 1981. The rest need to into "other foreign".
* The rest need to go into "other" but I can link those bits (roughly)

Getting "other" from 71/81 = fairly easy. Now I have to work out which ones are other in the rest - a tricky separation. Reminder:

* Not-other: Far East. We can put those in that category.
  + "Hong Kong, Malaysia, Singapore, Brunei"
  + Note: Can't get Brunei - subsumed in South-East Asia elsewhere. Slight error introduced (pop.~400K) but I'm not going to lose the other cats.
* Not-other: certain African countries.

..

From 71 far east: "and all Commonwealth Islands in the Pacific not elsewhere classified" is clearly this:

<https://en.wikipedia.org/wiki/British_Western_Pacific_Territories>

So that includes... nothing that's going to a separate country elsewhere. (Though note the handy "where are they now" "today part of" column on the right.)

..

Far East done, fairly easy. How about Africa? Oh wait -

* Other "New Commonwealth" not listed:
  + Mediterranean / Cyprus / Malta / Gozo / Gibraltar

Africa / new Commonwealth:

* I'm going to take my list of "new commonwealth" African from 1991 LBS. Thusly... Kenya through Sierra Leone.

<https://en.wikipedia.org/wiki/Decolonisation_of_Africa#Timeline>

2001 is looking fiddly: categories that cross over commonwealth/non-C Africa, I think. Let's look.

Oh, first: when 1991 is talking about "new Commonwealth", it's presumably using historical cats? (Well, obviously, if old commonwealth is canada/oz/NZ!) But let's confirm...

### UK Africa colonies (and dates of independence)

Sod it, let's be thorough! ->

* South Africa, Egypt: 1910/1922
* Libya (bit of a tangle, this one!): 1951
  + Emirate of Cyrenaica, part of Libya: 1949
* Sudan: 1956 (w / Egypt)
* Ghana. 1957
* Cameroon: 1957. But part-French so split.
* Somalia (nee Somali republic): British/Italian. 1960
  + Republic of Somaliland split from this in 1991
* Nigeria: 1960
* Sierra Leone: 1961
* Tanzania (Tanganyika): 1961

Then a big clog of British secessions from 1962-68:

* Uganda
* Kenya
* Zanzibar
* Malawi
* Zambia
* Gambia
* Rhodesia (now in Zimbabwe / Zambia)
* Botswana
* Lesotho
* Mauritius
* Swaziland

Then a gap until just:

* Seychelles, 1976.

..

1991: some of the "non-commonwealth" countries that were part-colonies are not included. It doesn't seem straightforwardly racist: Egypt is in there. Though... Why is South Africa excluded but there's still "old commonwealth" colonies that precede it by a century? Err.

..

So that's our new commonwealth African countries in full. Now let's check what 2001 does: cross those boundaries? (Looking in 2001 census definitions)

* North Africa: yes, it contains Sudan (and Egypt).
  + Where's Sudan in 91? Doesn't have own category... Ah: it's in "other Africa". Again, not a totally British colony.
  + So 2001 North Africa is **good for "other". All non-"new Commonwealth"**
* Central/Western:
  + we have in data -> Congo, Nigeria, Sierra Leone and "other".
  + Nigeria, Sierra Leone: **new commonwealth.**
  + Congo**: not**
  + Other? **Split.** Some are new commonwealth, some aren't.
  + OK, need to flag this up for later work. Possible fraction or summat
* South and Eastern:
  + Kenya, Somalia, South Africa, Zimabwe:
    - One not (S Afr) rest are, can divide. Then ->
  + Other: **split again**.

So: some **definitely commonwealth or not**. The "others" split. At some point I could allow functions into the re-code to, say, split that randomly or whatever the user wants (for monte carlo-ing).

* For now: **other probably needs to go into other** on the basis that most populous countries **do** have their own cats.

..

Africa 2011 now. What we got again?

* Nigeria, Kenya, Zimbabwe all with own cat - so into new commonwealth Africa
* South Africa not.
* I'll track down the exact details of africa "other" later. Needs a bigger piece of work to think through. For now, same - "rest of world"

..

Which leaves us... where?

### Bits

Some more new-commonwealth bits, right?

"New commonwealth other" for 01 and 11:

* Doh: some are in "non EU-countries in Western Europe" (Malta/Gozo). I'm gonna leave them where they are for now. But that means no more "new commonwealth other" for 2001.
* 2011? I think again "other non-EU countries" or "middle east". Will have to do for now. Checking impact of different re-assignment on results will be useful.

..

Which leaves "other" doesn't it?

Newp: **New-commonwealth** **America**. Needs splitting from the rest and sticking in new-commonweath-other. So what's that then? ->

* Also: Ceylon/Sri Lanka into other N-C.
  + **Uh oh: no Sri Lanka in 2011.** With these uncertain ones, need to come up with an uncertainty strategy. They can only go into one of two bins, so I can try various permutations. There's a paper there: showing where this uncertainty matters = important.
* New C America is?
  + "America — Barbados, Guyana, Jamaica. Trinidad and Tobago and other and unspecified countries in America (including Bahama Islands, Barbuda, Bermuda, British Honduras, Cayman Islands, Falkland Islands, Leeward Islands, Redonda, Turks and Caicos Islands, Windward Islands and West Indies (so stated)"
  + Via <http://www.onlineocr.net/> converting 71 definition screengrab. Nice.
* Which means for other decades?
  + Well in 71 and 81 that all goes in / is already in "other NC".
  + 91 has its own NC America "Caribbean" section. Plus Belize and Guyana, actually.
  + Full 91 list for NC America:
    - Barbados, Jamaica, Trinidad and Tobago, Other Independent States, Caribbean Dependent Territories, West Indies (so stated), Belize, Guyana
* I think we're then going to have the same problem of being split. Where '91 has its own separate non-NC Caribbean...
  + 2001: Jamaica / "other Caribbean, west indies".
    - So what's in "other Caribbean"? Can't be that many left. Check non NC-Caribbean in '91:
    - "Cuba, Dominican Republic, Guadaloupe, Haiti, Marie Galante, Island, Martinique, Netherlands Antilles, Puerto Rico, St. Barthelemy Island, St Croix Island, St Eustatius Island, St Martin Island, Virgin Islands (US), Windward Islands (Dutch), Windward Islands (French), Windward Islands (Portuguese)"
    - So again: red warning! But the solution is probably to check on relative levels of numbers, to see what the error rate is likely to be.
      * Compare 91 non-NC to NC Caribbean. Difference is? Just downloaded. That's 103 vs 292 (presuming non-NC central america doesn't come under the same heading)
        + **1630 for NC America vs 99.** So yeah - a reasonable assumption. We should do the same for our African diffs with this table. (Again, 1991 being very useful!)
        + Actually, let's check that Africa thing: NC vs non-NC in 1991: 58 vs 262

Namibia, for some reason, is in "other commonwealth" for 91. But ballpark figure:

7294 vs 6993. Huh: so **considerably more problematic** not being able to separate them. Though note, actually, I probably **can** separate a lot of them - the most populous.

* + - * What's non-UK vs UK in total for 91? (All minus UK, 1 - 10)
        + ~150K out of ~5 million or 2.98%. Err. What if we make "anyone not Scots" a foreigner? (1 - 16)
        + 444K Scots makes for 10.89% "foreign". Hah. How does this compare to England?
    - For now: **stick these into "other New Commonweath".** (Much less of a problem for now than Africa, see above.)

Hang on:

* **I can make a NC-America category after all**, can't I? I have the Caribbean numbers in all. OK then!

Done. So what's left on my New Commonwealth list? Anything? Newp: **the rest is "the rest".**

Done done! Next:

## Pulling five-census bits together

First job here:

* Re-run geography intersects to get re-codings to 91 postcode sectors.

I should move this code to the census github page. Into Python folder, right? Where is it currently? OK, moved.

Now, to make a little function. There's a common zone to intersect **to**. Then bunch of ones to intersect with it. Then a folder to save to.

..

Going well - just need to check if 2001 OAs need turning into unique features.

Err. Which involves. Err.

### Fun with postcode sectors

So all the shapefile data needs tidying, it turns out - apart from 2011 OAs.

Postcode sectors in particular: duplicates, with some being "Loch". How many aren't actual postcodes? Oh, just Lochs.

On combining: another approach might be to pick the largest area ID, since most of these are adding what are probably uninhabited rocks. But let's leave that for now.

..

OK, we now have all unique-ID per row files. I suspect postcode-matching is going to be imperfect! Except for those intersected cos they'll take the ID from the SHP itself. But I can get on and run all the intersects now. (And all that tidying code needs putting in its own file and docing.)

..

Right: intersect creation code running. Remind me what happens next? Think I need to hive off another R file - the one I'm using got a bit tangled I think.

Currently:

* runIntersectsAndDataThroughReassignFunction.R

Ended up the unique zone ID checks. That should be somewhere else. I need one doing **just** the re-assign work.

Oh - and before that, I need to save the common-cat CoB data. I think keep the 'common name re-assigns' in a separate folder from the original census data: keeping stuff together that's related to particular combos.

(3.6 minutes for 4 intersects. Nice. Probably need to check them all.)

..

Hmm. More issues. It was nearly working then. Problem:

* **Intersect is dropping some zones completely.**

My only guess as to why: there's some self-crossing polygon issue. I did have an R way of checking this. Wonder if I kept the link or code...? Yes, **intersectPolygons.R**.

http://gis.stackexchange.com/questions/163445/r-solution-for-topologyexception-input-geom-1-is-invalid-self-intersection-er

With this method for removing self-intersects (which could be in either dataset):

ed71 <- gBuffer(ed71, byid=TRUE, width=0)

..

So let's try that then re-run. Just running the Python... seems to be taking longer!

Newp. Let's try same with PCS... If this doesn't work, things are going to get fiddly! Other things to try are going to be... well, let's see first.

Newp. Let's try the original non-unique-ID'd files.

Newp. Same problem in the original data.

Let's see if 2011 IZs has the same issues... Not as badly (nowhere near) but still a little. Just confirmed it's def leaving holes. Next: try a GRASS clean-up, via:

<https://faunaliagis.wordpress.com/2013/08/14/bad-bad-polygon-fixing-invalid-geometries-with-quantum-gis/>

Then try R.

..

OK: QGIS menu intersect works fine. You know, I might just need to do this manually for now - need to get something done. So...

Hang on: need to run on the processed files. So... yup, all working fine so far (one or two missing from 71 but nothing at all major).

Well! I'll have them done anyway - hopefully they'll still work fine with the next processing stage, huh?

### Output census re-codes

Which will then need to go through my nice manually gathered intersects. Which will hopefully work! OK, saved here:

C:\Data\Census\VariableCoding\CountryOfBirth\_fiveCensusRecodes

..

Done two. Everything working. Cripes. Though note:

* If there's a clash recode of the original large-zone ID column, it can't then merge by that name. Solution would be to tell it what column to use in the large data, huh?

..

Now, there's not very much point in making more maps just yet I don't think.

..

Oh: I just accidentally did a cool thing:

* Prop.table: gives proportions - I thought initially per column, but it's done the whole thing.
* **Which is our total population figure**. Whatever that actually is, this is the population count in % terms from this year. They should be comparable. Nice!

..

We **can do** columns with it, turns out:

<https://stat.ethz.ch/pipermail/r-help/2007-June/133260.html>

prop.table(mymat, margin=2)\*100

..

Hang on - that **does not** do what I need. But it does tell me how I should do it:

* Just sum by zone and compare to that (which prop.table also does - margin = 1. Woo hoo!)

So actually the column margin gives us what Geoff wants: CoB as % of total CoB.

..

I think it might be worth kicking out some longit. Data graphs:

* General CoB over time - % of total pop. Or similar.
* **Change** over time. Which CoB cats changed the most? I could do with some geog way of showing the same thing too... my last GRIT graphs seem like something that might work:
  + E.g. whole graph is Glasgow region; x axis is time (four change points); y axis shows % change between censuses.

..

For which I'll need to run the rest of the data through the geog. Or in the case of '91, attach it - and I'm expecting some postcode fiddling to be done. Let's do 01 and 11 first!

OK, for morning:

* 1991 just needs data linking to zone.
* Hmm: % change graph would need same process of aggregation but to something like TTWAs. Would be better to do that from small-zones, obv. (Though can look at nesting.)

Before making graphs n shizzle, I should probably think about what else needs doing to keep things moving forward / what to say to Gwilym. For now, tis time to buy pasta-y things. And hopefully some Kale.

### 1991 data linking

So all I need here: 1991 postcodes and 1991 CoB data at postcode level. Oh, which I don't have, do I? Doh! I'll need to get that, then make sure... err, hang on, do I? Yes I do. OK then.

Note: those zero row gaps - probably the lochs etc. Though we'll need to check since we're doing this manually.

..

Fun with postcodes! This was never going to be neat, was it? Various issues - as well as different formatting lengths **within** the zone data, there are (probably) some errors like "PH35 4" and "PH354" almost certainly being the same postcode.

Thing is: the space is needed because outward postcodes (first part) are variable length. Any that are five-long - well, there aren't any.

..

A couple of other longer ones - but only PH35 4 that's ambiguous. And according to reverse geocoding, "PH35 4" is actually "PH32 4" (it's a tiny little zone). "PH354" is the larger one that's clearly "PH35 4".

We should check on things in the census data before starting to change anything...

### 1991 census PSEUDO postcode sectors?

That are fairly different from actual 1991 postcode sectors. Now we need to work out which of these two I should be linking 1991 census data to. I think that needs its own file. And it's time to start a test folder to fit with package stuff.

<http://rmflight.github.io/posts/2014/07/analyses_as_packages.html>

I appear to have read this before...

<http://www.nrscotland.gov.uk/files/geography/products/1991-census-bkgrd.pdf>

"In Scotland, the census equivalent of the ward is the postcode sector."

OK then! "These are not necessarily true postcode sectors as they may have been split to recognise local government district boundaries. Within each local government district, each pseudo postcode sector is assigned a two character code in the range AA to FG. For ease of reference, pseudo postcodes sectors are called postcode sectors."

Good lord, that's ridiculous! So it's probably going to be pseudo PC. So will have to do all the re-assignment/re-processing again. Ahahaha.

..

Various amounts of faff. Turns out zero-rows in CoB data for 1991 really are for specific zones. Just going to check the original data, make sure I haven't accidentally done something to it...

Newp. The population data equally has no-one in those places. Let's just double-check that.

..

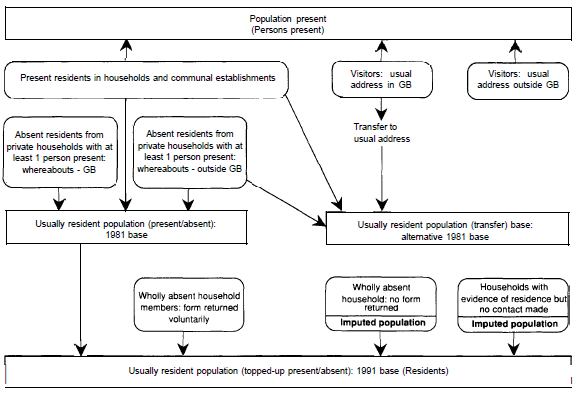
Well, this is all getting weirder. LBS vs SAS populations:

* Zero columns in LBS, seem to be those with <1000 in SAS.
* The top-level numbers...? I need to download again with sensible names!

Looking weird. Do we have the same problem with England? Also note: the current five-census coding: I think the small-zone 1991 coding actually maps exactly onto what I've already done (because we know which is colonial and which isn't.)

In the end: I may have to think about it later. Having realised that 1991 SAS is - for 5-census work anyway - precisely the same categorisation (I think), I should probably just do that and revert to 2011 IZs as target geography. But just going to check...

Oh, useful. From 1991 definitions:



Though I'll still want to use CoB totals I think.

..

### LBS ShmellSchmeeSchmess

So there's this (91 definitions):

* "The full Local Base Statistics (LBS) will be issued in England and Wales as abstracts for wards or sub-divisions of wards with both 1,000 or more residents and 320 or more resident households. Otherwise, the statistics for a ward will be amalgamated with those of a contiguous ward, generally following the procedures for EDs. The SAS will, however, be issued for wards which fall below the thresholds for LBS but which are above the minimum thresholds defined in paragraph 1,33 above." (5)

The SAS method: re-assign to a contiguous zone. Probably the smallest contiguous zone but not always.

And we appear to have no record of which zone that would be. I'm thinking there's **probably** a way to track that down from using the SAS zones - but note we face the same problem of assigning zero then moving it. (<50 people / 16 households).

In fact that should be fairly easy: total pop being re-assigned would obviously show up as a difference in those neighbouring areas.

..

Just going to have a look at Scottish districts. Four to choose from for 91! Let's try "scottish district wards".

We'll have to track down by ID type. Let's get the CASWEB district data, see what we're looking for.

* Census data: 57 districts, four-digit ID.
  + Not ward districts
  + Admin districts: wrong ID (2 digit) and slightly too many zones.
  + Postcode districts: no way enough wards. Err. Hang on...
  + Electoral districts: exactly the same as admin districts. Probably the smaller number of... oh screw it, I'm stopping for the day!

Actually, quick note: the entire point of getting districts was wrong-headed: I can only use the smaller zones to check where counts have been aggregated.

Reminder for morning: point of this - without finding out where/which zones those amalgamated numbers went, it'll be impossible to aggregate correctly. It'll involve a lot of manual arsing about.

### LBS workarounds

Two things:

* Turns out, for the 5 census harmonising: because of the colonial distinctions, 1991 SAS can map directly to the more detailed LBS categories with no lossiness.
  + But that means a different 1991 small-scale geog. So it's probably better to use 2011 IZs again (or possibly 2001 since those might be closer to 1991, but it probably doesn't make a huge amount of difference.)
  + So I should manually run those intersects since the Python code appears to be doing something different. (And I have to go back and work out what's going on there too!)
* For the 3-census harmonising (1991 to 2011):
  + This will involve a little work: I need to figure out **which LBS zones** the zero-count ones got re-assigned to. I think I can see how to do that using 1991 OAs to figure out which postcode sectors have much too large pops. It should, in theory, just pop directly out of those count results.
  + That will then give me a list of zones to re-assign. I guess I could also do that automatically if those count results tell me which - because we know it has to be a contiguous zone, even if we don't exactly know which one. So that's a fun little problem!

But I should probably get on with the 5-census stitching for now then return to thinking about "the Africa problem" (ahem).

### Re-doing 5-census to 2011 IZs

First thing: I haven't done the 1991 output area dissolve checks - unique IDs that is. Sooo.

Other thing: I'm not even sure the "unique" test is necessary. We find fractions by ID in the intersect anyway - doesn't that end up the same? I could check I suppose... Or not bother and just get this done. Another task for later.

On 2011:

* Scots data warehouse: no direct IZ download. So aggregating with intersect again seems the way forward. Unless... Ah, no. But there's probably a look-up isn't there? Yup:

<http://www.nrscotland.gov.uk/statistics-and-data/geography/our-products/census-datasets/2011-census/2011-indexes>

..

OA\_DZ\_IZ\_2011.xlsx in the Scotland 2011 census folder. Nice. (Assuming we have all IDs!)

Now though: 71 and 81 to 2011. Err. I did already do that, didn't I? Yup! Only one to do: 2001 to 2011. For which there's probably a better way of doing it than this but hey ho...

..

Let's see about this lookup. So:

* Take 2011 output areas. Merge in IZ code.
* Done.

So two more jobs:

* 1991 CoB data needs column-combining in a different way, based on the SAS data. Which I have, right? Yup. So needs coding up. Do in Excel sheet again? Yup, copied sheet across. Copy variables...
  + Noting that "South East Asia" in 1991 is the 3 ex-colonies.
  + OK, that was nice and easy. Now to re-code/re-run the re-assign...
* Just run the 2011 intermediate geog re-code through the column re-assign. So that's now done. Just need to geog-reassign the other four... (so again, the order of code is getting tangled!)

Dammit I have to leave now. Let's try and note down where we got to:

* I've output all the column-combines including for 91 SAS small scale zones.
* Next I have to put them through the geog intersect re-assign. After that, I should be able to do some quick prop.table stuff to get proportions and make some maps. Maybe. (Before tracking down other vars).

..

### Errrrrr

Remind me what I was doing? OK, I've read it. I think I remember...

**censusData\_smallToLargeGeogs\_5\_censuses\_to2011\_IGs.R**

Is where I'm putting it all together for those targetting 2011 IZs. 2011 is just a lookup/aggregate, which I've done in that doc.

..

Note:

* 1991 SAS census: **ID columns need first two digits removed and any white spaces removed.** The one remaining non-match is "Loch". This should be in a tidy-data section at some point, really.

..

Well: we appear to now have five stitched census CoBs, for the 14 categories we can currently use. (Though with some uncertainty esp. in Africa.) What next?

For maps/data, I want to turn that into CoB as % of each zone. I should also remind myself what other data is needed - each of those datasets needs to go through this same procedure. At this stage we probably don't need the actual census population. But what else was there...? I can't find it above. Let's look in the chapter again...

### Other variables

Oops:

* **Index of local house prices**.
  + How did I not see that one? Where on earth am I going to get that going back to 1971? Oh hang on - isn't it in the census as well? No. No, it's not. Idiot boy.

### General mulling \*salutes\*

So.

* I don't know of any house price data pre 1990. Or 95 in England/Wales. Uh oh. Didn't think this through very well!
  + See? <http://www.zoopla.co.uk/askme/details/house-prices/stockport/55211#sIRZ4tU5CsoSeEVg.97>
  + Halifax: <http://monevator.com/historical-uk-house-prices/>

There are other census data I need to get hold of. That all needs feeding through the same geog re-assign / lookup process. Which I guess could do with being made a little more streamlined huh? Given the time we have left? Anyway, they are (not counting population, which I'm just going to use CoB totals for):

* **Number of dwellings** via household count**.** (Then num dwellings per acre from that.) So just number of households - though note that's not necessarily a perfect 1:1 mapping, it's pretty good. "Living at the same address" is key to all the household definitions so it's not bad.
* **Local employment.** Which is going to be "economically active". Probably. Right?

And that's it. But that's a fair amount of data prep and faff there.

..

Going over the jobs I need to do again. Brain feeling a little bit better than it was yesterday. So what jobs do I need to do?

* Well, things a little tricky with house prices - there are none prior to '90. Which suggests the need for the 91/01/11 census stitch to get done pretty rapido (though worth talking about this on Friday.)
  + But for the data I've got, we need summary stats per IZ per time period. (6 months either side of census period?)
* CoB data needs converting to "share of migrants from that country" per zone. But that's just prop.table for columns so easily enough done.
* I'd also quite like sum rows to pick out some maps of % changes per zone.

Oh and probably:

* Linking all census data together into one time-period-indexed CSV (which will allow some nice ggplotting too.)

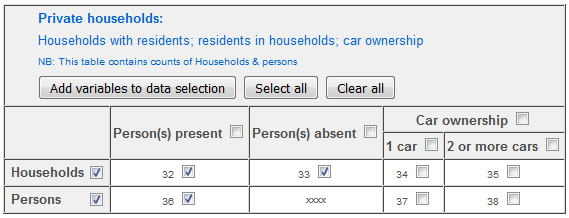
Then other data:

* Dwellings / area, then employment ("economically active") hopefully.
* Both of which then need feeding through the intersect re-assign.

### Other census data 1: households

Let's get these first. Some already downloaded for 2011 but I don't know the column names. So. Dwellings - given definition of "household", using household count.

* **71: SAS15, private households** ->



I \*think\* (judging by the numbers) that 32 is the number of private households. Let's just say it is, shall we? Err, but then what does "persons absent" mean? Let's press on to other years...

Just checking SAS19 "household composition and housing" that also has a total households figure... Yup, same number! Good good.

* **1981: SAS81 "tenure and amenities"**
  + Note: both here and in 71: zones with fewer than some threshold value (lower than 8 households is the minimum I've currently seen) are set to zero.
  + But this is at least confirming the same pattern for 71.

So that's interesting. The zero cells: potentially tricky. Given the aggregation, there's a fair chance they'll end up in the same zone. But... may need to do some tracking down of... err.

Actually, there's no smaller overlapping zone from which we can work out where they got re-assigned. So that's a bugger! Can't even interpolate because that'll be adding values. So. Yes. Bugger.

Larger zones possibly less of a problem, but... well, it **would** be possible to work out **if** the smaller zones were not on the boundary of a larger zone. You'd know... at least, that it was in that larger zone. Err. Does that help? Maybe not!

Hopefully it's less of a problem in city areas. We'll see.

..

Actually, there is a way to do this. It's an interesting little spatial problem:

* For those zones contiguous to a zero zone
  + Which are beyond a certain SD from what you'd expect given spatial autocorrelation? (Pulled from the others) The variation may be too random of course...
  + Then, if found, apply some rule to split that in a way that interpolates a sensible value.

..

* **1991: SAS 20 tenure and amenities**

Looks like we have all households here. This is Scots first OA year, the zones were, I think, shaped to have roughly equal numbers of people. Or households. Probably the former.

Note: it's '91 that will need the ID re-processing again (no spaces, drop first two digits).

...

* **2001: Errr**

Several options. Housing stock; dwellings; accommodation type (households) contains all households; amenities - contains "all occupied household spaces".

And I can guarantee they'll all be subtly different.

I should probably stick to "all households" for consistency. Which is in UV056 accommodation type...

From this we can also compare household and dwelling number. Thus, dwelling vs households, means then SD:

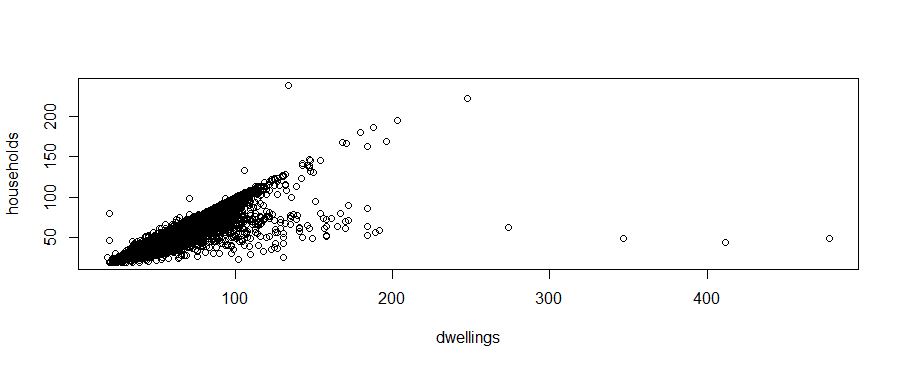
54.15700404 51.45634213

15.91146876 14.11857603

..

Not too dissimilar but (right way round) slightly less households than dwellings.

Though with some annoying diffs:



So some places with many more dwellings. (Really? Surveying number of dwellings in the census?)

So let's just go back and check whether household is **really** all we have in the earlier ones! Yeah, think so. Humph.

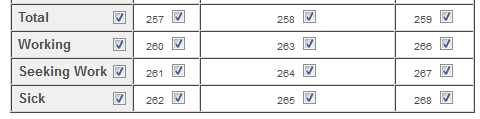
### 2. Employment

Which is "economically active" (EA) innit?

Oh good: no, it's not. You can be EA and unemployed. Apparently. Err. Hang on, didn't I just sum up 1981 numbers...? Urgh, let's look again!

* **1971: SAS05 - economically active**

Is going to require a little summing. Columns are: male (e.g. 257), then: female single/widowed/divorced (e.g.258) then female married (e.g. 259).



So: **total EA minus working gives us EA not working, for employment ratio.**

Interestingly, the **total** columns aren't always quite the sum of the others. Or sometimes larger. So that's nice. But total/working for 257:259/260:266 should be our employment figure.

Hmm: interesting point - how to normalise for changes in male/female employment patterns? Indeed, should you? This rather sexist table might offer ways to think about that. But. For now? Press on!

Well, one way would be "share of employment" again, wouldn't it?

..

Note also: 15 and over. 1981 is 16 and over.

Note note: "EA" includes self-emplloyment, according to the definitions (unlike the total EA count in 1981).

* **1981: economically active**

Think I can just get total persons EA and total not in employment. There's a separate self-employed cat - that should be EA but is it? Let's check using all totals.

Uh huh:

* 1981 EA figures - "Economically active" is a **separate cat** from "self-employed". Duh. And it's a big difference. We could do with knowing what that is in the others.

So to match 1971 **will need to add self-employment in**.

Though according to its definitions, as far as I can tell, self-employment is **meant** to count as EA. Huh.

Scratch all that:

* EA includes any economically active person, employed or unemployed. We want that ratio.
* The '81 table is a bit confusing but does make sense: we just need total persons EA vs total persons not in employment (also in EA, remember!)

..

* **1991: economically active**

08: economic position. And what I've got from this -

Males/Females separately (no available total)

EA (M/F - 12/166)

On a govt scheme (which I think should count towards unemployed) (M/F: 67/221)

Unemployed (M/F: 78/232)

..

* **2001: univariate table / UV028 / economic activity**

So in theory we just need "economically active" and "unemployed" (0002 and 0012).

Just noticing the small number of unemployed here, for OAs. They're bound to be randomised a fair amount. Close quite close to nowt may be problematic, %-wise. Might be worth **sticking em into bands of low/med/high or similar.**

### 2011: housing AND employment

I didn’t do this above. So in theory I have both tables - just need to work out which.

KS105SC: Household composition. This has all households.

KS401SC: Dwellings. OK then. Can see how different like with 2001. (If consistent across zones, may be useful.)

..

KS601SC: economically active. Bonza.

### Tidying that shizzle

So got all that data but it's not very neat or quite in the right format. Sooo time for a tidy up. Then after tidy-up, stick it all through the appropriate geog re-assigner.

**household\_n\_employmentCensusDataTidy.R**

..

So I've just done housing - which was the quick one. Hmmph. Now for unemployment, which will be something of a column-rearranging/staring at casweb faff.

..

OK, that was a fairly intense day's data processing. Housing/employment done - though housing with some issues of course. But done for now.

In theory, I should now do the housing data. I might just do that in the morning while writing things up as they are. I guess all the datasets need combining too huh? Well: probably best to wait until I hear exactly what form the data should be in. Let's go home!

### Getting housing data

Point in polygon then price averaging for a given time window. No faffing, just the same 2011 IZ geog for all.

Oh wow:

<http://www.inside-r.org/packages/cran/sp/docs/aggregate.Spatial>

According to google, I've already read that, but...

<http://www.maths.lancs.ac.uk/~rowlings/Teaching/UseR2012/cheatsheet.html>

One line! A while to run, but not stupid long given the number of points:

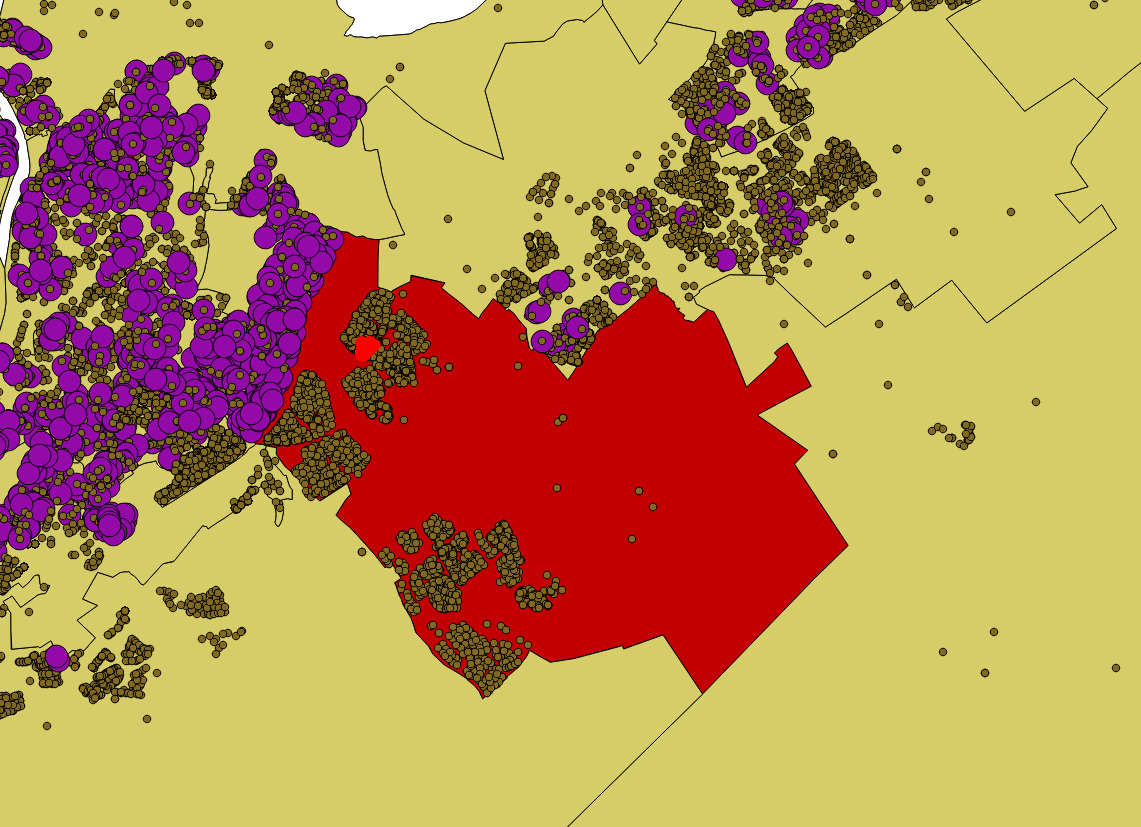
hse@data$interzone <- hse %over% izs %>% dplyr::select(interzone)

..

Suspicious: large purple circles are sales before 1992. (I have zero values for 1991 census in that, and two other, zones. That seems highly unlikely doesn't it!)

One of them's Harris, so maybe. But...

Note: this is from the raw merged data. Huh. Starting to think that was new-build after a certain date. We shall see...



So let's carry on with the assumption that those are probably new-build estates. Could be entirely wrong, but let's see.

..

OK, meeting done. Let's just drop the current data to Geoff. I can then go back and re-run for higher-res categories - though that'll also require footling about with neighbouring postcode sectors to deal with zero-zones.

For now...

<https://cran.r-project.org/web/packages/spdep/vignettes/nb_igraph.html>

Just doing a spatial weights matrix. Wondering about Islands but might send this one for now.

Compare, straight queen contiguity and k nearest neighbours (4 in this case)

Which I did in **miscDataReshaping.R**, by the way:

|  |  |
| --- | --- |
|  |  |

### Some more checkinz

From Geoff:

Thanks for the data files; I’ve had a look through them all and think I understand. Obviously since all the locations are coded, I don’t actually know where they are. I wondered whether the file Scotland\_ig\_2011.prj might help but I can’t open it.

On one small point, the number of observations for the price file (1276) seem to be fewer than for the other variables (1279) unless I have missed something.

The numbers of country of births is smaller than in my original paper, but this is to be expected given the finer spatial scale and I don’t think it will matter too much.

..

First, what did I send him? A shapefile presumably? Oh - somehow, lord knows how, it lost some of the key files in there. I should also give him a csv version of the same file, perhaps with centroids.

Then: I did make some notes about what's going on with the housing years...

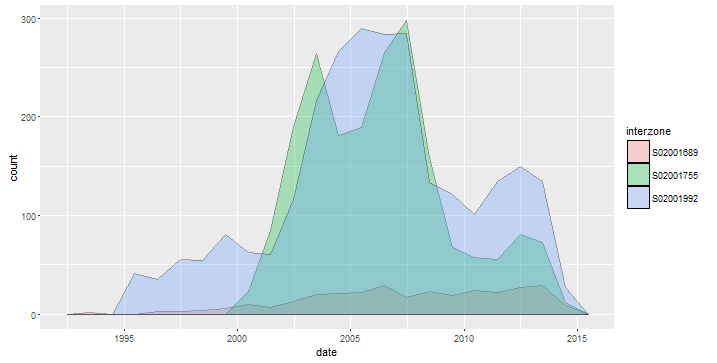
..

Well that's curious: the exact same code is now no longer working. Uh huh. Wow... so finally got it working again. That was a pain. Checked it's right. Now back to checking what I wanted to in the first place...

OK, found one thing out:

* The 199 obs (some the same house I think) that didn't get zone: they're generally very remote, and generally slightly incorrectly geocoded so they're just over an edge. So it's **not these** that are missing from the missing zones. That's the next thing to check.

OK, sorted:



Two didn't have housing in that earlier period (though one does have some from mid-90s).

## Updating for 1991 to 2011

After which Geoff says he can run the estimations. Two things need to happen here. 2.5 things:

* Work out where the zero postcode zone counts got re-assigned to. Combine those units. (2001 and 2011 will then be re-assigned by intersect.)
* Do the re-assign for **both** levels of CoB categories.
  + So that's the other job: need to finalise that harmonisation across CoBs for the larger number of categories.

For which I will need tea to begin.

..

### 1991 pseudo-postcodes: working out where to re-assign

So what's this going to require then? I'm just checking on the various shapefiles vs the data files. I think I've got the pop data files but need to double-check.

Urgh - feel like going back to square one with this just to check I'm getting everything correctly.

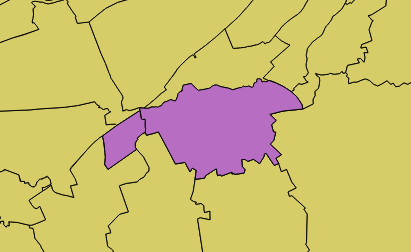
..

I've got the one-ID-per-row pseudoPC file. And each shp zone has data. Only two left over, one of which is shipping, the other one's probably... thing. You know. Other thing.

I also now need to do the same thing for pop for 91 OAs, to then do the re-assignment checking.

..

Done some checking. Just need to think through what I'm actually then aiming to do. So staring at this, where the two purples are PCSs with zero-set data:



That's two things to consider:

* Neighbouring zero-set PCSs, so working out where continguous values went could be tricky. (There's others with denser contiguous zero values.)

But the first thing to do is the counts. And what's the theory again? Remind me?

* Say that PCS on the right has 1540 recorded pop. If I re-assign OA pop counts to that PCS and it's very different from that, we know it's been given the count from one of the zero-set.

So I need to do that population re-assign first. Which requires the intersect geography be created first... which needs doing manually. Let's put it in:

**C:\Data\Census\Intersects\Scots\_checking91\_PCS\_oaIntersect**

They perfectly tesselate so should be pretty spot-on...

..

Hum: it turns out there's a label naming convention that should allow me to bin the data from OAs into PCSs.

BTW this is all happening in:

**Census/tests/1991\_pseudoPostcodeSectorChecking.R**

So there's just some aggregating to do. Let's just check on variable names again... I just need to use one of them, as they're directly comparable. Sooo.

* s010064: residents present 1991, sums the first four. Can use that for comparison.

OK, it's working. I'm stopping now but this is where things are getting left:

* pop91a: this contains the values from both side by side, and you can see where the re-assigning is happening. It's exact and tends to be neighbours: it should be possible to identify exactly which PCS got the extra numbers.
* Then of course we're making the assumption that's the same assignment happens to all. That's probably something worth checking by looking at e.g. CoB numbers using the same procedure.

..

Right-ee-ho, so now I'd like to... oh, note this via Richard / quant-geog list:

<https://www.liverpool.ac.uk/geography-and-planning/research/popchange/outputs-and-events/>

Just registered for workshop, 14th July. Some overlap huh? Stil, pressing on with my stuff, obv. So

..

So let's look at these numbers and remind myself what I need to be doing. I'm also just trying to work out a decent way of working across two or more machines, including git and dropbox (really don't want to have to sync over git each time I use different machines...)

Symlinks (or windows' version of) might be a solution, though I'm... hmm, probably not: dropbox may end up syncing symbolic links as actual files.

..

I could do with not worrying about that and getting on with the thing I should be doing. So let's look at it again.

* 91 OAs: aggregated by the PCS code that was in the file itself (handy) giving nice exact numbers.
* Obvious for a lot where the re-assignment has happened, but we need to do two things:
  + Method for working out which of those it is
  + Method for those that contain more than one re-assignment

And need to remind myself what the goal is here:

* Some postcode sectors have zero values, cos they're below a threshold value. Their counts get re-assigned to neighbours.
* I want to identify **which** neighbours using the OA counts. I can then **combine those PCSs into one** - larger zones but correct values.
* For some, there's going to be multiple re-assignment...

So looking at the data in R, a plan forms:

* Pick the columns with zero values from the PCSs.

Got that list of values that can show us where the re-assign happened. Now I need to identify exactly where it went. Which we'll do like this:



* 770 is the OA value, showing us what that zone **would** have been if not set to zero.
* The next row got those values. But we want to be able to detect that by searching. Which we can do with:
  + Does this row's PCS count = sum of its OA count and the zero column's OA count? If yes, mark it with the zone it needs to be combined with.

We'll see how many we get that way. The next stage might be easier done separately. There probably isn't a quick way in R to do the above. Or... err. Might be?

Hang on, let's think about this. Probably need to do some kind of list apply to the list I've just got of zero zones...

Except it's not a list. But I can index it with a list! Err.

OK, started making attempt, but not going to get that done tonight. Time for homewardnessness.

..

Allo! So - I was being dumb. Figured this out as soon as I walked away:

* The total count for all non-zero zones? Substract the OA count for that zone.
* Depending on if that figure's unique (highly likely but we'll see; if not, we'll need a continguity check I suppose...) can then:
  + Merge in the other list! Bingo!

And then see about the remaining. Hopefully small enough number to track down manually, but we'll see.

Oh: they're not unique. I can tell that from the first list. So this now suggests the need for a contiguity check as well. Which is hurting my head a little.

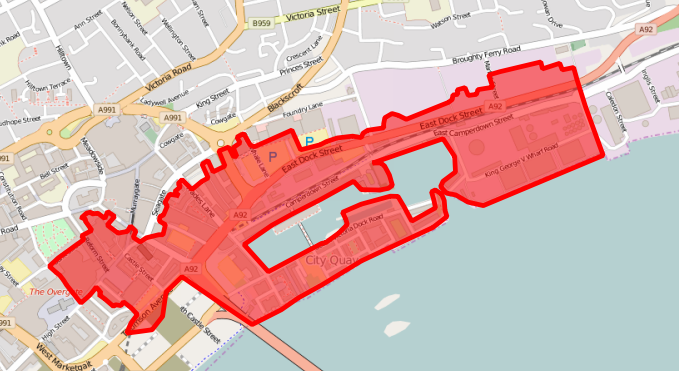
Let's look at doing that... how did it work again?

**miscDataReshaping.R**

..

### What are the non-matching PCS zones?

In the pseudo-PCS data. So this is in the shapefile ->



Label "w". Dundee port. Is the only one of the 1002 zones that don't match the 1003 in the data we have. One of those in **the data** is "shipping", but the other one might be this, mis-labelled in the shapefile. Let's check...

Yup. DD1 3 is the postcode, that's the right area. So need to replace the label in the shapefile...

OK, fixed that. Lord, I'm dragging this out! Next up: the continguity matrix, that I'm not quite sure how I'm going to use yet. I can't see a way to stop it being loopyloopy. Don't really want to give up for the day cos TIME IS RUNNING OUT but think I may have to.

..

So really really need to get this done now! I think sticking it in excel to look and work out the best approach is probably the thing to do.

**91\_pseudoPCS\_joinTesting.xlsx**

So let's see. I've added a little pretend contiguity matrix. So the thing we're after:

* A subset that reduces the minus check to contiguous zones only.
* The only way I can currently imagine doing it:
  + Zone label in place of each contiguity binary.
  + Concatanate them all into one column
  + When doing the match check, do a grepl for the label name

Before I try that (given there's no guarantee it's actually going to not still find some duplicate zones) I should see if I can reduce it, see what's left, what it looks like.

I also need to think about the next job here: **I need to use this information to aggregate other 91PCS tables**. i.e. which zones to add together. I suppose get this job done first!

..

I think I need to look at this stuff in a map...

Hmm: the vast majority of the remainder, I think are those unique values where more than one got assigned to another zone. Which is where it gets tricky - and it's either some quite complex algorithm to hunt them down, or by eye. Let's just look if it's obvious...

Hmm: mapping is messy. This is actually quite complex, which might explain why it's fucking with my head.

One thing though:

* Any old/new count that **has a difference** shows us which zones we should be targeting, doesn't it?

OK, I have a map that's starting to work: can see where I have matches and where don't but need some. Those ones - they're smaller in number, so it must be they require multiple inputs. Err. There may be a way to do this in stages with matrices... But let's look first.

OK, so going to try the combination approach for the last ones. There's one little error but I'll come back to that one. The tricky part here is getting the zone ID refs from the combos (probably).

Given R does vector combos, can use that in apply to index a dataframe?

..

Well I have my combos, but there's a problem:

* Before, I was just subtracting

Oh hang on, no there isn't...? Oh yes, there is!

* Previously: the OA value subtracted from the PCS value **would match** those that could match.
* However, those with **multiple OA value inputs** now won't. There's nothing to directly match against. It's slightly fiddlier... I need to...

Hmm, this might have been easier in Java. Let's take this example:

PCS: 2152

OA: 1755

Difference: 397

That 397 value - it's coming from a combbb... oh Lord, I'm back round again, isn't that what I just made??

The thing:

* The OA value: **that's what it should be**. So we know the difference **is what's come in from possibly multiple others.** So we should be able to get some matches...

Well, we have some. Some possible issues. And note: other add counts go up to four zones. It did then get added to the next zone in the list, which is going to be something worth checking...

Food, then a closer look.

Pairs worked well. Others didn't but we don't have many left now, so looking in map and seeing what's what.

It would be useful to mark the zero-zones we found matches for...

..

Well, that was very nearly almost going perfectly. I've found some where the zone label don't match. Mainly in the Islands. I just completely cannot imagine a reason why that could have happened - but it's going to mean re-tracing over the whole thing, because we need to make sure identified labels are correct. I think actually they probably are in the data itself.

But something's gone wrong somewhere - so tomorrow is a fun day of carefully going through it working out WHAT THE ACTUAL FUCK.

### Fixing the shapefile

Or whatever the problem is. Going to piece together everything that's been done so far. The problem **shouldn't** be in the data file - the re-assignment itself is happening without reference to anything geographical.

So I'm going to start with a fresh file and piece it all together again, see what happened.

**findWhereZeroCountsWent\_91pseudoPCSs.R**

previous file:

**tests/1991\_pseudoPostcodeSectorChecking.R**

First thing to do, check through the original sectors and make sure everything is where it should be, before tracking down where any changes happened.

An example of the differing label to work out what happened? That was...

6756AE (in the original, I believe). Scalpay off Harris.

..

* Good: IDs to single rows is fine. Problem appears later than that.
* Version saved with Lochs removed and Dundee wrong label zone fixed: also fine.

Actually, this shouldn't be that hard. It **has** to be when the data's merged back in, doesn't it?

Ah, good news folks - it's something pretty simple:

* When merging shapefiles, don't merge shapefile@data, do it on the main spatialpolygons dataframe.

OK, now I have to remake the QGIS file for seeing what I need to see. Which is what, again? Err. Let's put it back together from scratch.

**V2\_checking1991pseudoPostcodes.qgs**

..

OK, everything put back together! Didn't need to redo everything, just one small error, phew. But now whole new QGIS to look at... what's left?

..

Urgh. I'm getting sorely tempted to Java this up now. This is totes an optimisation problem.

OK, just for the sake of argument, imagine I was going to do this. What would it look like? Or just work through what's left and think about it, might be the better option. They're mostly right, I think...

So:

* Pick out pairs with problems:
  + So this is for zones that got two zones assigned to them. (And only those where the single zone didn't find something; I'm defaulting to those, though should probably check...)
  + Actually, there's only one that's problematic, the original one I found: 6332AT, got 6332BL and 6341AE. It's surrounded by a bunch of others, all very islandy. I know I found another 4-group that matched easily, so I might just make some additional groupings then see what's left. But remove this one that's most probably weird/wrong.

..

Going OK: possible problem with zero re-assigns file misidentifying... oh no there isn't, it's just that one from the pair that got wrongly picked up.

Onto the final re-assigns now. Seeing it in map form, would have been impossible without it...

..

Good lord, done. Jesus, that was a mission. Err. And there's more to do yet.

### Phew! What now?

Two things:

* Dissolve the polygons that need dissolving so that other decades' data can then be intersected with it.
* Aggregate any '91 data with those new categories. Will have to remind myself what that data looks like / which datasheets we're using.

Success! Blimey, this is such a palaver. I haven't even got to the CoB recoding yet.

Brain death may have set in. May need to give up. If I don't have it quite, I don't.

..

### Data aggregation

Step 2 of above: converting existing data for 91 LBS to the new polygons. So need to sum all the counts. I could also do with just checking that the same reassignment zones do, in fact, apply. I might do that first, actually.

Just checking: ah. I think I see what's happening here -

* Numbers are all over the place. Why? **Because small-zone OAs get re-assigned too**. So the lower levels are a bit awry. Not so with pop counts, they're all over the limit. None of those will have got set to zero, though, I don't think, so you can't tell which ones... No, think they're just randomised.
* The same zones are zero at the larger scale and... hmm. Well, we'll see.

Let's look at another table pairing, something we need from both. We had OAs before, so e.g. economic position. Let's look at that one.

Huh: tables have slightly different structure; different col names for the same things. Plus need to read my... ah no, I just need to be comparing like with like. Sooo.

* So see "economic position" above. Having looked: same basic categories, but SAS/OA level breaks down into M/F so will need summing (didn't I do that somewhere?) whereas LBS/PCS won't
* So let's just compare "total persons economically active" in both. Actually, while I'm downloading, might as well get all the fields I need.

..

I've emailed the data service to see about this whole thing. But for now, I need to press on with the assumption that the zone reassignments are consistent across tables.

### Econ active

So let's carry on with the same thing: using economic activity. Have a little think about what needs to happen.

* 2001 and 2011 data itself doesn't need to change - it just needs re-assigning to the new 91 geography.
* 1991 data should just need aggregating.

..

**91\_PSCzeroCountZones\_aggregateTables.R**

## Working towards superdiversity conference

With all of 8 days to do it in. First thing - this from the 5-census data email to Geoff:

* That's three census variables in separate CSVs for each census: country of birth, household count and 'economically active', which contains total 'economically active', total unemployed, and percent 'economically active'. There's no separate total population: I've just been using the sum of CoB per zone to avoid differences to pop counts. (Though accurate population comparison across censuses is a particularly thorny issue.)

Percent 'economically active?' Should be percent employed, shouldn't it? 'Economically active' is the total number of people we're considering in that category. So need to check that's what it should be.

I could also do with reminding myself of all the steps in the process. In fact, making some slides from that would be good, so need to do it anyway. Currently I'm fairly confused by it all (and it's not really in a format that others could very easily use.)

..

I'm also keeping an eye on this:

<https://www.liverpool.ac.uk/geography-and-planning/research/popchange/introduction/>

So, a full SDAI grant looking at doing the the same thing. Not quite the same: I'm aiming for larger geographies, not any kind of interpolation etc.

A paper related to some of the methods Chris Lloyd (PI) uses:

Lloyd, Christopher D. ‘Are Spatial Inequalities Growing? The Scale of Population Concentrations in England and Wales’. Environment and Planning A, 23 December 2015, 0308518X15621306. doi:10.1177/0308518X15621306.

### Brainstorming the process

What are all the bits and how do they fit together? Just to remind myself.

* The final aim: a table containing rows for each geog zone, a column for each census.
* The things you need to do to get there:
  + Find a common geography.
  + Make sure census table columns are also common. This is **not** the case for **country of birth**.

And how did that end up looking in code/folders? What have I got, what is it all doing? Let's work this out while thinking about how I repeat it for the 1991 re-coding.

* [CoB\_columnNameUpdate.R](https://github.com/DanOlner/dataStitching/blob/master/CoB_columnNameUpdate.R)
  + Sorting out the CoB columns. I'll need to do that for...? I think I may already have done it for 1991 postcode sectors, by the look of it.
* [CoB\_stitchingCategories.R](https://github.com/DanOlner/dataStitching/blob/master/CoB_stitchingCategories.R)
  + Again: I think this is already loading all the correct data. I just need to apply the coding for the 3 later censuses.
  + Re. 91 recoding for the larger PCSs (i.e. summing some rows where COB-stitching is summing some columns): I don't it matters which I do first, just as long as I know I've done both.
* [censusData\_smallToLargeGeogs\_5\_censuses](https://github.com/DanOlner/dataStitching/blob/master/censusData_smallToLargeGeogs_5_censuses_to2011_IGs.R)\_to11IZs.R
  + Takes pre-formed intersects, produces the re-aggregated data. Err, apparently only for CoB. What happened to the others? Oh: that's done here ->
* [household\_n\_employmentCensusDataTidy.R](https://github.com/DanOlner/dataStitching/blob/master/household_n_employmentCensusDataTidy.R)
  + The other variables. Not sure how much that will need editing to get it to the new re-assigns.

Obviously, I could definitely be more systematic about all that. Except - the data is messy, so it's not that straightforward. Anyway, need to start somewhere with this (and work out where I've actually got to).

### Quick check on 5-census employment data

Specifically: did I use "economically active" correctly? Err, yes I did. Phew. That's nice and easy then.

OK, so what's next to get this period done? Shall we do the hard bit first?

### New CoB categories for 91-11

Let's just look at the last run and see what I did. Probably have to go through it in the same detail.

OK, let's just go through it! Remove all colours, start again...? Actually, probably useful to have the other one open to see the rough blocks.

I'll keep some of the other colour blocks. Or possibly all now I look...

Just editting in a new R file:

**LBS\_3Census\_CoB\_stitchingCategories.R**

(Random thought: higher-grained may not be better since we lose relations between country groups. May have to think about other forms of proximity... oh, you could probably do err, something like a Mantel test: is there a relation between spatial proximity between both? I have a feeling someone's done that.

But there may well be some pretty graphs... : plot spatial proximity of countries vs Scots proximity, over time. Hum.

..

And so far just going through and doing the recodes. Ones that may need to change:

* "Rest of UK". Are there any larger breakdowns now? Yup, done.
* Europe: 2011 is the limiter here. There's "members other" / "accession other". So let's code up the members to 2011 first...

There are a lot to match. Is there an automation process? How long would that take?

Not too long, it turns out. Now I have to work out what's left to match.

OK, for the rest:

* **Other europe:** just chuck the remainder into that bucket for both. There's maybe some eastern/other split but ... huh. The 2011 definitions doc does not seem very useful for this. The CoB classifications there are large / don't match / have no notes. Which is going to make it tricky working out some of the breakdowns...

..

Mostly quite a lot of throwing into bins. Stuck again on the various Asias. Let's see what goes where.

OK, I \*think\* all the remaining ones are going to have to go in "rest of world". No good categories left.

p.s. all those other yurps - they need to go rest of world dude. Turns out that's what I did with the 5-census ones, which the colours actually don't make very clear.

* I could do with tabulating how many from each group we're talking about and comparing. In fact, that'd be very useful for Scotland through those censuses - overall change graphs. Yup!

..

Huh: categorisations that I made before, I'm less convinced of. Just going to have to go with them for now. Having the other larger categories helps.

OK, we have a first pass at combining CoBs. It worked first time, which is nice.

..

Next? Currently CoB and the other vars are done in different files. (In fact: employment / housing in one then house prices in another I think.)

..

Well, I need to just press on with this. OK, job 1:

### 91 PCS zero-counts removed: aggregate data to the new zones

Which I've already started on in:

**91\_PSCzeroCountZones\_aggregateTables.R**

I've got most of 'economically active' done. I just need to get it matching the previous formats. And the previous formats were...

Unempl EA prcntEm

With the last worked out from the first two.

OK, onto housey stuff for 91. Previous section was "[Other census data 1: households](#_Other_census_data)". I should check whether from 91 on we have a dwellings count or not... But what did I have before?

Ah so: only 1971 didn't have dwellings. That's helpful.

OK:

* 1991 LBS60: dwellings and household spaces. Some summing to be done but it's all columns, so...

Arse: **households have different zero zones counts than individuals**.

Yeah, so. Best option: for household number for 91 - aggregate from smaller geographies as we did before, but to the new PCSs.

### Intersecting

That's all the straight matching I can do. Now I need to intersect:

* 2001 and 2011.
* But also 1991 dwellings cos the zero-counts are different.

Then get house prices. Then we're getting there.

**3census\_LBS\_dataIntersects.R**

Is the file I'll use, based on the previous structure. Err... need the intersects first, including 91 OAs to the new PCS.

Huh: dwellings at OA level has (unsurprisingly) the same pattern. What's odd though: **households don't**. We have those values.

Which isn't great, looking at the difference between household and dwelling again... not a good match.

I'm tempted to just leave off density for now. I could harvest average house size from somewhere else but... hmm.

Let's just check that other "dwellings" tables have the same issue... well, they're going to aren't they? Can't imagine why they wouldn't.

OK, one possible simple approach to dwellings (but still more faff):

* Adjacency matrix; then, average dwelling size in surrounding areas.

Huh. Note:

* Some of the dwelling counts are things like 2 to start with. That's too small for it to be a threshold issue. I might stick to the averaging process anyway, but... odd.

Nearest **seven** neighbours is the minimum to avoid NAs.

..

Well that was... possibly useful code at some point, but actually a waste of time. Wrong table. Should have used:

* **Occupancy of dwellings and household spaces [LBS65]**

I think that table has the zeroes in the right place... let's see... Yup. Well that was easier!

### And intersecting again

What's left? Well first - have to get dwelling info for 01 and 11. I'll have 11 already, just need to find the table (and tidy it). 01 will need downloading.

I have to wonder whether that's a priority. Density overall... well, per postcode area, I'm not sure it's really density, street-wise. We have a measure of per-room occupancy that might be better.

..

Actually: after all that, I reckon let's not worry about other housing stats just yet. I should have enough just with employment and house prices.

..

CoB done. Now for employment. Err. Can I use the same data I have? I can, can't I? OK, done. Relatively painless. All the functions still work - yay! Nothing untoward, looks OK on a map. Sad I don't have it for all the UK.

Which leaves

### House prices

Which I can just aggregate by zone, again using mostly similar code to what I've already used, and with updated prices. When I'm done that, I can leave!

OK, first pass done on housing data. Tomorrow should be "get this all together" into some story.

..

**Housing/1991to2001\_twoYearBandMeanPrices\_1991\_PCSnoZeroes.rds**

A lot more data shaping tomorrow but at least it might be in pursuit of something to actually show. Time to leave!

## Findings!

Not, like, a model or anything. Probably. But it is time, yea verily, to attempt looking at the numbers we have. Whatever they are.

I wonder what numbers? Well, the obvious ones first:

* Top-level Scots CoB change over time. (I should work out something about rel to ethnicity really...)

So I have some basic top-level things. Just a toe in the door. A long way to go to produce something useful - so let's have a think of what I'd like to aim at, so I'm not flapping around too much.

Obviously, this is all for Scotland (which, it turns out, is probably much less interesting). But - things from the abstract:

* How has diversity changed in the past four decades in the UK?
  + So some simple measure of diversity change. Of which, of course, there are several. I'm not really big on single-number measures
* And what causes some migrant communities to remain spatially concentrated and others to disperse?
  + Stability! Anyway. This is possibly the tricky one. How to show how certain groups stay put or don't?
  + In fact... **do some** remain concentrated and others disperse? There's also the fact that we can't actually directly tell this from CoB. But we can look at the dynamics of movement and assume...
  + I'm thinking of some nearest neighbour analysis over time. If I could get **just** that done, I would feel like I had something.
* A model that explores the interactions between migration, employment and housing...
  + I am unsure what to do about the employment / housing thing. Could do some simple correlations.
* We will present an approach to stitching consistent census geographies together over the time period covered, building on methods pioneered by Walford, Norman, Rees and others.
  + The easy bit! Need to remember to do it though.

The rest, I'm not so sure I'm going to worry about just now:

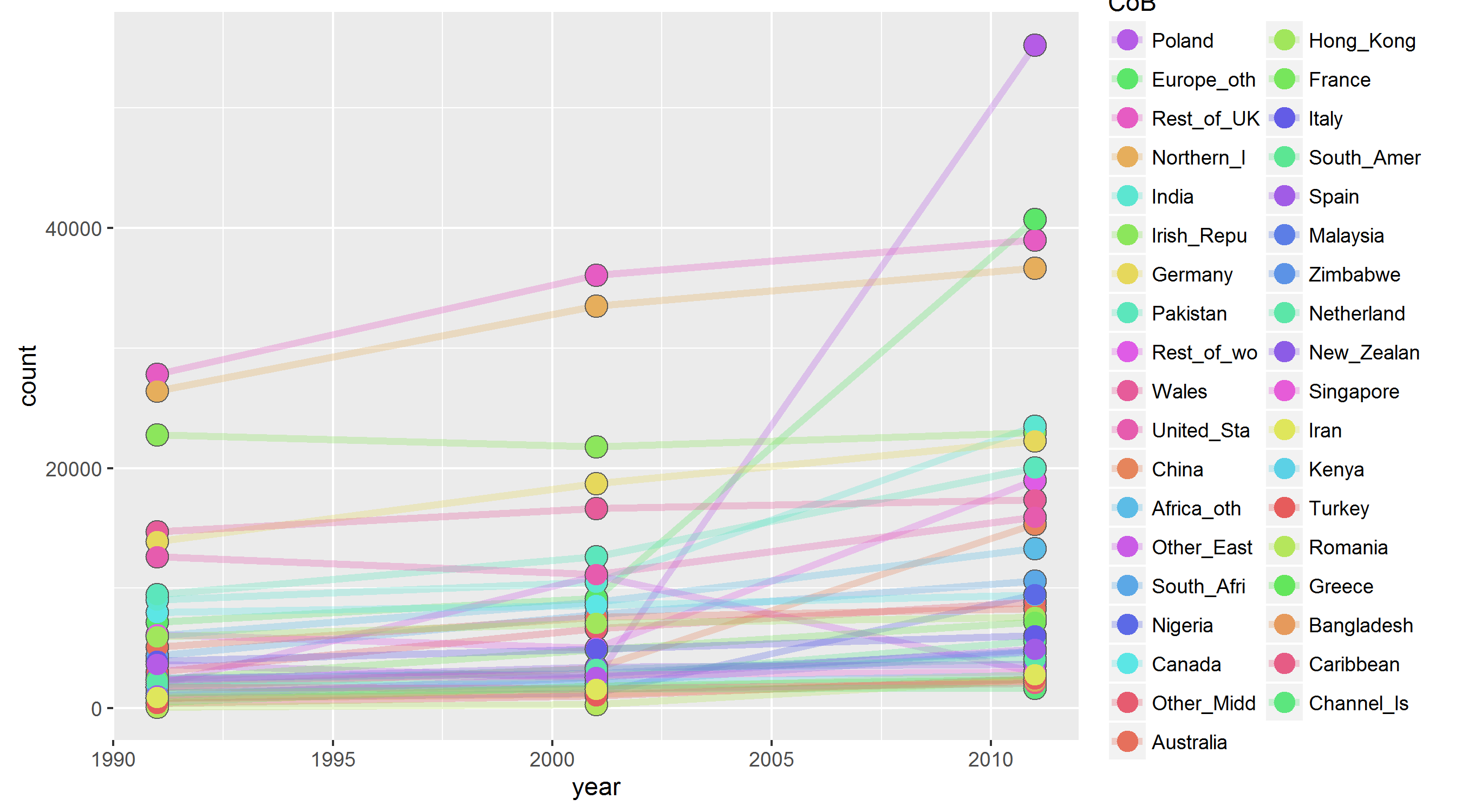
* We will also aim to explore more dynamic questions, such as: what are the long-term impacts on house prices, market fragmentation/dynamics and economic agglomeration? We will do this by extending the econometric approach set out by Meen (2013) where house prices and employment are assumed as given - i.e. they are included in the model as potentially important 'exogenous' push/pull factors for migrants. However, house prices and employment could themselves be profoundly affected by the arrival of migrants. We will address this with a novel econometric framework to estimate the feedback effects on these critical variables (i.e. that will allow local house prices and employment to be “endogenous”).

So what first? Can think about going back to the 'five segregation measures' paper.

### Some basic plotting

Too many categories to make much useful sense of the outputs. (Note other midd.1 is still in there, though I've removed it - must be remaining factor.)

Ah, this one is semi-useful, after having worked out how to reorder:



So ordered to the max value of each, picks out the top ones a little more clearly.

Let's get some Scots/Eng/Rest of World graphs out first. (Tots and proportions.)

Done that, with the usual levels of ggplot fafferooness. Now what to do for the rest? I could do with a function if I'm going to be constantly repeating this stuff for subgroups. Though probably I want more control than that...?

Quick and dirty's probably the way forward at this point! Sooo.

* How to say which were largest groups in the different decades?

And! What else am I after? I need to start thinking about the spatial stuff. If I could function that shit up to get something useful out, that'd be good. What did I say above again?

### Correlations n shizzle

Let's try some. With employment levels in those zones to start with perhaps. For which I'll need the data.

OK, so I've got employment loaded for all 5-censuses. Aaand now what?

* I want to know if particular CoB groups are generally going to less well off or better off areas to live, over the five censuses.
* So I'm interested in: **for each country group, where is it most prominent?** So for that CoB group, what's its proportions?
* Then pick out some top number of IZs where they're mostly going.
* Then correlate... err. What am I correlating? If I were doing three-census, could do house price vs employment with CoB as colour.

So actually let's try it with 3-census first. That was surprisingly painless. Now I need to colour some of those by CoB across those censuses. So need to do some picking out of CoB groups - which to assign to which zone? Many will be in the same zones...

Yes. Many will be in same zones. So need to think about how I show that. Errr.

Different ways:

* One: Break down by country, then order each country's zone %s. Then facet by country while sizing based on % (or % rank, actually).

Yeah, let's start with that. Can at least get at look.

Right, so started to get a look. There's more in there somewhere... More tomorrow! I have to think about **change** and how to show it. And the background change: how did PCSs generally shift in their positions? In the general shift to more employment and house-prices...

..

Morning! So there's still this to do. Had some thoughts on getting stuff out... wonder what they were? Must also remember that I have to make some slides. I should get on with that in the moments when my brain's a bit flurby. Some thoughts:

* Plotting locations... some clustering measure based on simple distance and how that changes. But to start with, just plotting (with some jittering possibly to avoid overlap).

..

OK, I've got a nice little thing showing how employment vs housing has shifted about over the three censuses.

### Spatial

Let's do some other stuff just to poke into some spatial patterns... Hum. OK, so I've got something. What do I want to do with it?

### Change!

Can I pick out some different groups with particular features? Such as:

* Which ones shifted socio-econ positions the most over the decades? i.e. moved up the housing/EA curve?
* Which CoBs tend to cluster in certain wealth bands? Which are more spread?

Just thinking about accounting for proper house prices:

<http://webarchive.nationalarchives.gov.uk/20160105160709/http://www.ons.gov.uk/ons/publications/re-reference-tables.html?edition=tcm%3A77-283149>

## Faculty conference

Err. So I've got this notion of getting the actual model working for a week Tuesday. I appear to have currently lost the fecking doc in the move. Err. So. Let's bimble through it all in an attempt to remember what the actual fuck. (And whether I'm missing any key data.)

..

Just a quick look at a calendar to work out what I've got left, time-wise, for this: 3 days (including today) with various levels of faff in them, then 5 days clear through to the conference.

Next job: remind myself where I got to with this. Let's look through the last presentation first...

..

Huh: did way more graphing of stuff than I remember. Particularly like the house-price / employment one. Sooo... I'm looking to replicate the actual model? In the time we have? Let's look at the paper again.

And here's the last email from Geoff (14th July! Good lord!)

..

*I think we should now proceed to estimation as soon as possible on the 3-census basis as a start; I think having the extra C-o-Bs is a big advantage and we will be able to tell quite a lot about stability from these alone. As far as I can see you have all you need to do this since the spatial weights matrices have been constructed. The 20 year lag should be sufficient to avoid some of the endogeneity issues. On the data issues, at least for the 3-censuses, I think we should use the housing stock as it can differ from the number of households (although some countries assume they are the same I think). You are right about employment (and I found this in my results); because commuting is so great at small scales, it doesn’t tell us much about employment in those areas. But it can be useful in a full model where we want to determine house prices as well as migration (or want to instrument house prices).*

..

I need somewhere to put the regression learning. Err. Let's do it in econometrics learning / "R Companion to Applied Regression / Fox".

OK, bored of basic learning now. Let's look at Geoff's thing and see what's what.

### Geoff's model!

Slowly digging my way into using vectors of variables. Random links ahoy!

<http://stackoverflow.com/questions/3585631/vector-of-variable-names-in-r>

..

As I'm learning about the sp-econs, some thoughts on visualisation occur to me.

* The spatial weights matrix: now I see what it does (the normalising / averaging version actually a potentially interesting way of visualising too. Just trying to think about what it means. The LeSage book does a Moran scatterplot - easy enough in principle, but I'm struggling to see how to interpret usefully.

<https://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#statug_variogram_a0000000601.htm>

Actually, I'm not sure where that gets us with migration patterns. We want to know... what do we want to know?

* Share of migrants.
  + Which has the issue of dilution. But let's mull that as we go along.
* So: "How much is share of migrants a function of average number of migrants in area?"
  + Note that normal contiguity matrix excludes "here", which would seem to be important. Is that in there? Yup! First term.

..

Just been doing some staring: it's coming together. Geoff's weights matrix is indeed yer standard contiguity matrix. But note that there's a **different matrix** for... err. Let's see about this!

### And the data!

While I'm trying to get my head around all that (happening slowly but getting there) I should probably get back on and remind myself what data I've got. The presentation was useful but let's double-check.

OK, so I'm still missing a few bits of the data: dwelling number, specifically, for a couple of the 3C batch. (Have to use household count as proxy for 5-census). But we have everything for yer basic stability checks. And I think that's what I'm mulling... perhaps trying to visualise rather better.

But! Using some contiguity matrix to do this. All that's allowing me to do: ask, what's the relationship between the average migrant proportions **nearby** (in the previous time period) and those in this zone now?

I'm not sure what to think about what exactly that will mean (and whether I should include a 1 along the diagonal) but we could at least look at it.

..

OK, just for the night, let's have a go at making the contiguity matrix and making the average / checking it's doing what I think it is.

Looking good - but just realised I've been doing raw numbers. So need to look at proportions tomorrow. I think there's a solid seam of visualising in here somewhere...

### And more on the data

So getting something from the weights matrices. An \*apparent\* very strong clustering between pretty much all CoB groups (on the basis of 8 nearest neighbours - could try a larger number and see?)

<https://beckmw.wordpress.com/tag/variogram/>

Looks interesting. I'm just trying to think about how you'd look at a range of different cluster sizes. This might not be the approach - I'm mulling raking over a bunch of contiguity matrices.

..

Right, so variogram: it'll report variance between all **pairs** of points and graph against distance.

<http://www.ats.ucla.edu/stat/r/faq/variogram.htm>

So we could do it via actual centroid distance? Which makes me think: there's an issue there with simple neighbours, isn't there? Doesn't account for distance.

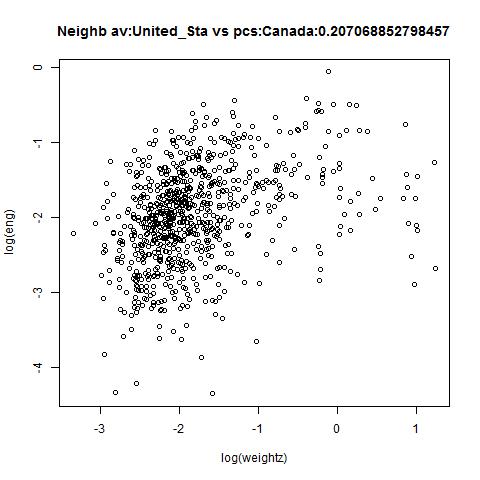
And the patterns: the very consistent patterns... well, let's test it by just comparing to random other migrant groups. I suspect the pattern it's picking out is just urban vs rural, but let's look.

I'm also starting to suspect a spurious correlation when using proportions like this. Esp. against places with large majorities in - will always be negative correlation to other groups? Well, that's only for random vs picks isn't it?

Trying to work out if proportions across Scotland makes sense rather than in PCSs.

### Randoms

This is a curious thing:



Strong clustering, but what does it mean? Clustering of the same low values together? Whu? A mystery!

<http://blog.minitab.com/blog/adventures-in-statistics/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit>

Just trying to hook onto something that will let me assess all this. And relatively quickly...

..

Err. Gonna need some restatement of assumptions given the time. What are we aiming for? I'd just spent yonks on the regression, now I'm playing with..

Hum. Variogram is gonna find all zero zones being close together. Not sure that's what I'm after. It's also going to pick up on correlations between cities - generally, I'm not sure I like it!

..

### Time?

Soo. Swinging back to the regressions and thinking about the comparison over time. I'd like to see what it looks like - which can include graphing some of it, but it'd be good to get regression numbers out.

So what we want? Well, picking on one set, let's translate:

* The share of migrants from CoB i in zone j at time t (latest census, or prev if we're looking at how it changes) is a function of:
* The same share in a previous time period, in the same zones
* The weighted matrix - i.e. the averaged values of surrounding areas for the same migrant group - in a previous time period
  + Also the sum of all other migrant groups from that previous time period.

So I can do one of those easily enough. Would be nice to get it set up to do a bunch. But let's pick one.

I'd like to know, though: what nature of the nearest-neighbour averages is **for the same time period** (which is what I've been looking at). If it changes... what is that telling us? Just thinking of those pics I've got for expanding, I think, Pakistani zones - how do you pick that out from this (if at all?)

Anyhoo: let's just look at some blummin correlations to start with...

..

Done it but those correlations don't look right (previous correlations do). Did something go wrong somewhere?

Note to self: I should make sure the weights matrix is oriented correctly when using x nearest neighbour. It's not symmetrical but who knows if I'm using it the right way...

Oops: just got date wrong. Silly person. Now I'm staring at them... what do we think? Does it get me anywhere?

### Where we at?

Time getting a bit short.

<http://stackoverflow.com/questions/14238148/how-to-load-packages-automatically-when-opening-a-project-in-rstudio>

Random link on getting things to load.

Now this!

<http://stats.stackexchange.com/questions/152958/how-to-avoid-log0-term-in-regression>

Man, why does no-one answer the question! It's not unreasonable...

<https://www.researchgate.net/post/Log_transformation_of_values_that_include_0_zero_for_statistical_analyses2>..

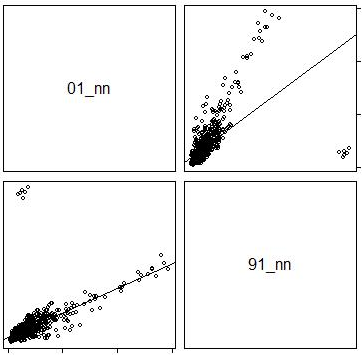
Oh, well, looks complicated then! Or better to leave as zeroes and not tranform...

### Neighbour averages: comparing weights matrices over time

This seems interesting to me and gives me something for the presentation. I've got a spiel working up: here's the regression equation. Here's the weights matrix. Here's all it's doing - nothing radical, it's just an average of some chosen set of neighbours. So it's asking: how is the value of migration in this zone influenced by surrounding zones?

But I also think there's something interesting in comparing weights matrices across decades. So e.g. here's what happens if neighbour averages **don't** change: actually, no, I don't need to put that. It's obviously just a perfect diagonal line. But what does it mean in each direction?

Random interesting point:



Cos OLS is getting its least squares on the vertical (I presume this is why) these two are radically different. Which seems pretty perverse to me, actually.

<http://stats.stackexchange.com/questions/20553/effect-of-switching-response-and-explanatory-variable-in-simple-linear-regressio>

Second answer has the pics - yup, OLS being vertical. Now in this case, neither way round is necessarily correct. And I'm wondering, even if I did have some reason to think one was really dependent... Hum: in this case it is - there's autocorrelation over time in one direction. So 91 "causes" 01.

[update: but what's the difference in the coefficient?]

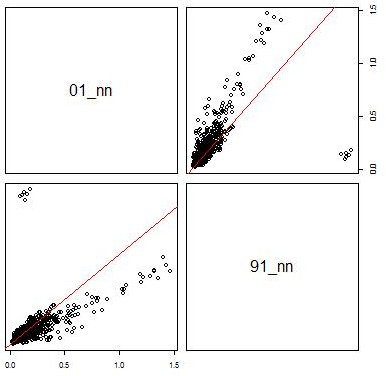
Anyway! My point was how it differs from the diagonal, not the regression line. What's the diagonal?

Ah: "line of identity"

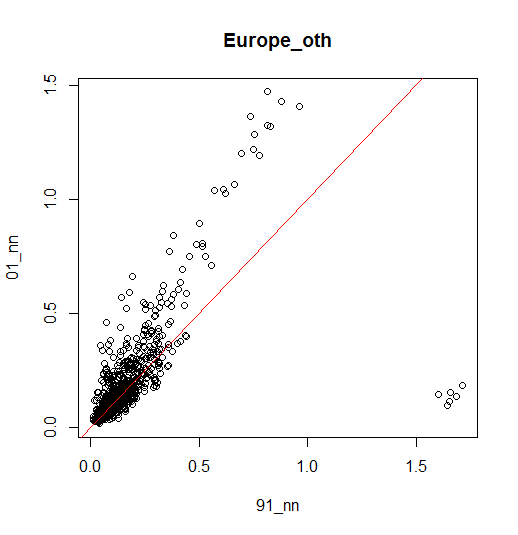
<http://stackoverflow.com/questions/17793690/adding-line-of-identity-to-correlation-plots-using-pairs-command-in-r>

..

Staring at this for Europe-other (ps. I think non-log is probably better):



What we got?

So havin' it so "01 is dependent on 91". We can see what looks like an early cluster (though they may not be spatial! Will have to think about that) that then disappears.

On the other hand: a bunch of clustered areas have increased their % between decades (above the line, easy enough to interpret).

While there's no direct spatial information, there **is** some here - **clusters** of Europe-other have been increasing in %zone size.

Err. I can capture that increase can't I? That might be a thing to do.

Note again: not loggin = better. The higher % values show up, which is what we want really.

Overall: seeing the pattern we'd expect if Scots numbers are dropping - most others are increasing. And - ah, this is a question - are the neighbour clusters with higher %s **increasing quicker than those with lower?** That gives us a stability measure, or a clustering measure. Possibly. Will mull that. Sounds very useful.

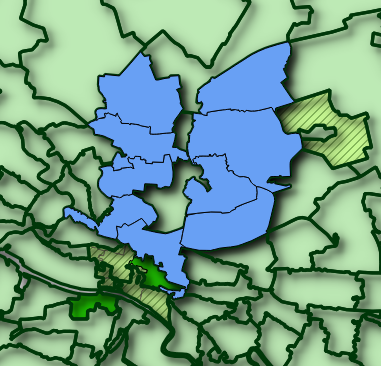
Though gotta think about scale: that people are being pulled into cities wouldn't seem unreasonable. Need other ways of breaking this shizzle down.

..

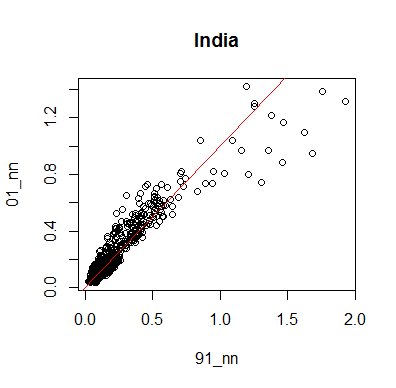
* Blimey: very **consistent German increase** across the board. I suspect this will be a "not spatial stability" case.
* Same for Greece. Even more marked. How to compare two?
* Iran: concentrated community in 6341BU. That's in Glasgow centre, pretty much. It's not there in 1991: but there is a bunch of high-ish values elsewhere in 1991. And... some evidence of clustering. Using 8NN causes it to be a lower value (I should check into this a bit - though it seems OK at picking this stuff out).
  + And there's one particular group of values where there's an increase in clustering (01nn vs 91nn). Are they in the same place?

Yup! So that's via picking on the weights shift between 91 and 01. So this is a new cluster. What does it do between the next two censuses? Question for later, connected to "how to show that".

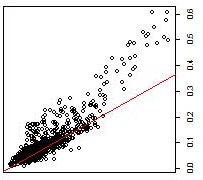
Update: see below. All this is doing: weights are **having their averages raised by that one zone.**



* India! ->

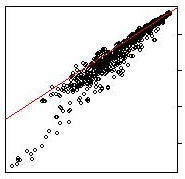


* Irish republic: quite large % drops in the higher values. Same case as India, moving to wealthier areas, or is something else going on?
* This little ramp at the bottom of italian weight vs weight:

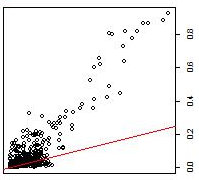


Is probably the marker between city and not? That's where a diversity index might come in handy (or rather more easily: how many Scots?)

* Kenya shows: absolutely no bloody pattern whatsoever. Huh. So what counts as "no stability"? (Very very small %s though).
* New Zealand: concentrating. Or, well... more in places with larger numbers. Cities again, right?
* N Ire looks like an interesting case. Again: need some metric of difference.
* Other East: loads more everywhere I think.
* Same with Other Mid.
* Pakistan: nothing much shows up in the actual numbers, but I know for a fact there's a cluster growth in Glasgow. So how would you pick that out? Err.
* 91-01: Polish numbers seem to drop? Again, compare to actual maps. Hum.
* Scottish pattern is clear: concentration drops in places where they're already lowest (going to be cities. Hey, I can colour code these by place or category!):



* Spanish: seem to have become much more clustered (0.2 to 0.8% av at the top; 1% in highest zones). Interesting that even only small %s can show such strong patterns.

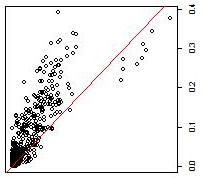


* U.S. population drops in certain places: military?

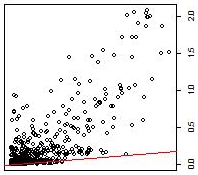
### 2001-2011

Some quite different dynamics.

* Iran: I think is doing the India thing of de-concentrating / shifting up in wealth. Needs checking.



* Italy the reverse: more concentration I think.
* New Zealand: would be curious to know if that's the same pattern...
* Nigeria: huge increases and a very lovely example of those increases happening in places already with high %s:



* Other east: huge absolute drops.
* Pakistan: clustering now showing up I think.
* Poland! Christ! Poland! Not a single zone where they didn't increase in number and concentration!
* Romania not far behind but much smaller numbers
* Scotland: more consistently below the line in all zones: across the board lower % clustering.
* Singapore: striking cluster wasn't there before. Where is it? Will come back to that! Mantel test? Distance between these points correlating to physical distance? No, that won't quite work in this case...
* South America: defo clustering.
* Spain: same.
* Turkey: same, but with an exception - a group (geog?) with de-concentration.
* U.S. clustering

### Wiffle

So the main story: v different dynamics across nationalities. Looking for reasons...

..

A number of the larger clustered areas have seen a drop in concentration (though gains elsewhere, in places with lower %s). **Now - is this a measure of dispersion?** Or not a measure: is it a signature graph? If I stuck a loess through that?

I think you could combine this with the LeSage book's quadrant approach.

Note: need some caution. Share values: could have gone down due to growth from other groups or movement from that group. We don't know. Other places that grew: could be new arrivals or people moving in from that CoB or others moving **out**.

So some caution, based on not knowing the mechanisms... In this India case, logically it makes sense (if I'm reading it right) that we're seeing internal movement: it has that kind of fingerprint. And we could look for some underlying mechanism: moving to wealthier areas?

Ah, we can do that! Housing costs, employment levels... Q: are CoB Indians moving to wealthier areas?

(And how can I show that change over multiple censuses?)

..

Need to remember: red line don't mean naffin really for... or does it? Hum. Hang on.

The "greater or less than mean" thing is worth mulling too. What would one generally expect for the average of neighbouring zones?

For England (for example): there's quite an even spread either side between zone% and neighbour%. Which goes back to question: **how to compare so clustering shows up?** In this case, there's the point about the spread.

..

Thinking about the quadrant thing again. Might not be much use between timesteps. But for start:

* If we use the median, so we're splitting "lowest half of zones" from "highest half".

Yeah, you know what: I don't that's very informative. Better things I could be working on. Let's see...

### Weights not?

Oh well, this is an annoying thing:

* Those spots where comparing weights across years: **they're not detecting vanishing clusters.** E.g. for the Europe\_other one: all that's happening - those zones around a **single zone with high %** end up with much higher average percent than other zones.
  + So all that's doing is **detecting zones near to another one with high values**.

Sooo:

* I can see how you'd only want the weights to be independent. In the case of that one Europe\_other value,
* And: that outlier **was** already visible. We just now know to expect - if it is an outlier - for the weights to show a clustering around it.

OK, putting that aside for a minute, let's look back at what we're after. This is making me think we might want to go back to quadrants again. Some way to link "high here and neigbours high too", and how that changes together / doesn't. I just need to understand summat about how those quadrants can best be picked. I'd like something a little less arbitrary that can really pick out clustering/stability.

You know, this might not work for anything. If all the weights vs weights is doing is picking out a focus on some outliers...?

Err. If I take weights **again**? What have I then got? Might be a bit pointless mulling that huh? Naw, I think I'm gonna need that connection between the two to say much of any use. So let's think about that. I don't think doing it arbitrarily is useful. It needs to start with:

* For a particular CoB, what's overall relationship between %s and its nn average?
* We've got the four quadrants but if we're thinking about change, what are we after?

..

Let's just look at a Mantel test type approach. I'm curious. Err, no. That's not going to work just with... and, no, there's no point. I think we decided this before: it won't work where zero values are obviously going to dominate.

Nearly time to leave... not the way I wanted to end the day!

Thoughts for later: how would you distinguish the other\_europe single zone high from those with high values around it? How would you then distinguish **change** in those?

### Weights yes!

Must be a way. Fresh brain, let's look again. I should also just check: it's oriented the right way, isn't it? Should be easy enough. Let's look... Yup, that's all good.

OK, so what to do with these weights things? For a start I could just run the bloody regressions. Perhaps try that to start with!

..

After some enormous faff dealing with slightly brok rstudio, we have some regression numbers... wondering what they mean. It's between time-points, so comparing to the identity line.

Which makes sense in this case since we're comparing how %s per zone shift about.

So there might be an argument for setting the intercept to zero? Remind me how I do that... OK, easy: y ~ 0 + x + z.

So I **think**  there's an argument, with shares like this, for just looking at how the line **changes angle as it goes through zero**. (And fits with comparing to identity line, which makes perfect sense to me for %s and comparing two time points).

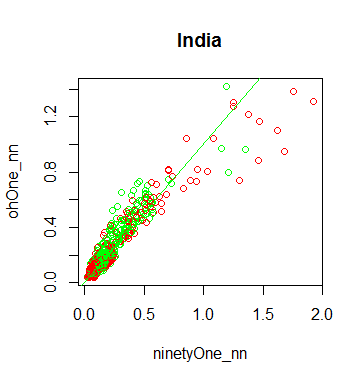
But would **not** make sense for the various econ variables, of course. But I'm not 100% sure why they're going in here at the same time since there appears to be no interaction going on.

So, simple things: **larger number, steeper slope!** This gives us a way to quantify the identity plots I've just been doing - though not, actually, to separate out what's e.g. going on with India, if we want to look for that difference at different scales. That would require ... something else.

Though with the India one... some caution due to previous issues! Need to think about that.

Though what I could do just to test the underlying idea: graph those Indian values against zone wealth proxy.

### Testing movement to wealthier areas: India



With red being "% employment is less than median". So it's mixed, needs more digging. But some suggestion that **most** 91 large av zones were in lower employment areas in the prev. decade.

Update: was using wrong employment year. That's even better now (used '91... should I use 01 though? As in, where they're attracted to?)

What about for property prices?

Yeah, not up to much.

..

Another thought while I'm thinking of it: it might be useful doing % of CoB over all zones (e.g. for India, sums to 100% over all zones - is telling us something about spatial spread). So the **ranking** of how that changes could tell us something. Combined with the distance of how those rankings change too... although ranks could stay the same.

Yeah, actually, it would make more sense to know how actual proportions change: if they spread out, we have spread! And possibly geog spread too.

### Idea! Underdetermination of movement/arrival

So just looking at the India graph above. When an area's average neighbour count is dropping (or just a zone), we don't know if that's due to movement or arrival or departure. Or: internal or external movement, might be a better way of putting that. Internal, we might know some of the forces, and something about the geography. Some of the same forces at work for external.

So the idea: spread your guesses for that, get a full range of possibilities with some underlying assumptions about the forces involved. Possibly constrain by other values.

But that's for later. Let's get back on the thing. Soooo.

### We need to decide what we're doing

For there is not much time! Err. What does the abstract say again? Is there anything there I want to aim for?

"How has diversity changed in the past four decades in the UK? And what causes some migrant communities to remain spatially clustered and others to shift? We will present a model that explores the interactions between migration, employment and housing for all major cities in the UK using data from multiple decades (1981 to 2011) of the UK Census."

I had a chance to tone this shit down before I submitted it! Oh well. The rest of it's all "wooo! Extending Geoff's model into endogenousness!" Which I definitely won't be doing. Ten minutes, not much I can get in there anyway.

I should also be mulling what to do with the 5 census vars also. Perhaps. Might not have time for that.

OK, so what have we got and what might we do with it? Err. Done some, re-run for continuity (which for migration will work pretty much wherever). Still not much closer to knowing what I'm planning to do. Think I might need to do some R mapping so I can get a better picture of what we're doing.

Things to mull while I walk to the shop:

* For the equation/regression, what do the values tell us?
  + For the t-1 "here" comparison it's a straightforward increase/decrease.

Hur. Just re-reading Geoff's model again: "share of migrants from CoB I" **does** seem to suggest it's **cross-Scots** %, not within-zone %. Which did make sense to me at one point: it's about where migrants are moving/drawn to/pushed away from.

Fuck knows. Let's go the shop and think about it.

..

Oh, a random thought on how to deal with Islands:

* Contiguity matrix returns zero for Islands.
* Nearest neigbour does not, and gives you a list. For islands, you can't have an exact contiguity number anyway, so you could just pick a number - then substitute in the island NN values for those zeroes.

### And try again with the what we're doing thing

Going to run out of time pretty soon. I may have to resort to scribbling on paper! So let's try and apply some logic to this. Point one:

* Weights matrices should, possibly, only be applied to / compared the zone in question, not to each other over time. That doesn't produce such obvious patterns but those patterns don't mean very much - they're most likely just picking up where single high-value zones are. **By themselves,** not much use.
* But can I use them in some interesting way, quickly-ish, when comparing to population change? Can I pick an example to stare at? The... hmm, no, those places with high percents out in the sticks, **tiny** pop numbers I suspect.

Oh, that reminds me, reason not to use cross-country proportions (CoB sums to 100% across all zones): each zones proportion is going to **also** be proportional to its size/population.

Right, so, back to this. One question is: should I filter out the smaller places so as not to have that bias? Well, let's have a look at pop numbers. Well, that's pointless right now! (Having done it.) I should probably stop trying to do pointless things at this juncture.

..

### Try again with the trying again

Couple of random links while I'm googling about the place:

<http://stats.stackexchange.com/questions/25069/2-period-difference-in-differences-fixed-effects-versus-ols>

One comment links to this:

<http://www.nber.org/papers/w20342>

None of this shit is helping me right now! ->

<http://stats.stackexchange.com/questions/118394/linear-regression-or-mixed-effects-models-for-data-with-two-time-points>

..

So the one thing I think I was after: the model! Or some info in relation to it. So stuff like:

* Using the spatial weights matrix to do some interesting work. Yes: this was my picture, wasn't it? Explaining what W is doing - just averaging %s.
* But then, what am I going to do with it? Again, I need to look at some examples, but this is at least a slight focus. (And my point about W needing to connect to the zone it's related to, to be actually meaningful. That might be wrong, but...)

So what have we got again? Remind me. I need to think through some possible metrics, something to graph to show where we're at. Err. So brainstorm then. Hmm. First brainstorm didn't help.

* We've got CoBs changing between censuses: largely going up while Scotland proportions drop.
* I've been after some measure of how those changes connect to clustering or spreading. So we've got a few things here...
* And I think the important thing's going to be comparing between CoBs. E.g. what's the general spread of the relationship...? Where do some fit in against that?
  + So this would be the same approach as with turbine/house prices: finding the general pattern, placing each CoB in the context of that pattern.

Actually... difference in difference, maybe. So:

* Have CoB proportions increased in places where they're already high?

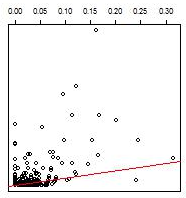
Err. Not quite DiD? Man, brain is fuzzy this weekend. Remind me what we were doing with prices! Oh, duh: it's in comparison to a control group. Although perhaps that's not completely crazy given I want some relative answers. But what would this look like? Errm. C'mon, this is not actually very complicated!

That was the right question:

* "Have CoB proportions increased in places where they're already high?"
* But also, have they increased in places where neighbour average is higher? And how does this differ between CoBs?
* And is there any CoB cross-correlation? Or summat?

So how the shizzle do I do any of that? I can't seem to mentally settle on anything.

Let's do some more staring at example. Greece:



That's with max 2% on vertical for 2001: big increases over 1991. Those zones with the larger initial values def are the ones increasing (mostly)...

Why did I say mantel-test stylee didn't work? Oh: zero values are proximate. That's not what we want. Or not entirely.

Putting aside the direct zone comparison for a mo, what about the clustering question? What connection can I make between zones and their surrounding averages? I've got a notion of using 4/8/16/32 weights to get a tail-off of clustering. But starting with the simplest, what would we be looking for?

* Could look at e.g. 2011 zones vs averages. So in the same time period, what do they look like?
  + And actually, visualisation-wise, this makes more sense to me. Should be able to then just compare change between censuses, **not** make zone% dep on t-1 weights.
* So I could start just by graphing this for a bunch of stuff. Probably ggplot. But graphing what? (Just to get a toe-hold on something I can work with.)
* Well: zone% in group1. Group2 is 4nn etc.
  + But lower zone%s won't show up if looking at them directly. I could normalise them to zero, but...

### Multi-weights-matrix ting

Hell, let's try that and look. Give me something to do! A couple more weights matrices then.

..

This is getting complicated. Codey faff hell. Where am I at with it? Got five different weight results worked out i.e. for all CoBs, av NN values over those five.

..

Ah: this seems to be exactly what I've been wondering:

<http://stackoverflow.com/questions/10357768/plotting-lines-and-the-group-aesthetic-in-ggplot2>

"Why is the group aesthetic necessary when the x-axis variable is a factor but the group aesthetic is not needed when the x-axis variable is numeric?"

Though I've read the answer a few times and still don't understand.

# AQMeN / England / Wales

## England/wales!

So a pretty hideously small amount of time. The one thing I'd like: English/Welsh data. This might not be quick, but could perhaps be quicker than before. I'm going to have to spend a little time reminding myself what I've got, where it all is, etc. It's going to be a faff.

The trickiest part, I suspect: working out the continguous value re-assignment for the LBS tables. A LOT MORE zones - quite a lot of them I ended up having to do manually. And I don't think I really have time to write a full optimisation program to do this. (Judgement on which is quicker...)

I'm presuming (might not be!) that all the CoB cats are going to be the same.

1991/2001 should be the same (and previous decades)... I have a suspicion 2011 might be different. If it does that... there will be blood.

Well let's what I've already got for England/Wales. Well, nothing for Wales it turns out - separate download. Fine!

### LBS rejigging 4 Eng/Wales

So two main jobs here:

* Make the new geography with some merged zones.
* For 1991 LBS, do some reassignment of values to those new zones.

The zones are wards, by the way. They were pseudo-postcode sectors in Scotland. Worth looking at size diff.

Now, I can't remember where I got the data for comparison. There was something... best look!

### Looking back...

Section: [1991 pseudo-postcodes: working out where to re-assign](#_1991_pseudo-postcodes:_working)

So R: **Census/tests/1991\_pseudoPostcodeSectorChecking.R**

..

OK, so 91 OAs: for england and wales, do they have the **ward code**, in the same way that Scotland had the PCS code? Then we can do some summing for working out what went where. Huh - beginning to remember what a faff this was.

OK, so let's get some OA data. Basic pop numbers again. What did I say I used? (Note I should be able to download Eng/Wales together though I'll prob need to stitch the shapefiles.)

I'm **really** glad I wrote down those notes so clearly! But let's come back to this with a fresh brain.

..

So it was population table **s010064**. Let's look - OAs and Wards, England/Wales combined. Actually, it's EDs for Eng/Wales at this time... hopefully they'll still tesselate the same. We'll see.

Yup, 1991 base total pop. Getting for wards.

Is there again a straightforward label naming convention that links nested zones? Are they even perfectly nested as they were in Scotland? Let's see!

England: yes. Wales? Yup.

**Note: have to use Welsh generalised Wards - 1991 full polygons have missing zones.**

Good. So now:

* Need to stitch eng/wales geogs.
* Work out if we have that easy way to nest them from the names or not. Err.

Now of course it would be good if I could function this up, but there's probably not time...

Yeah, we may have a new problem: a lot more zeros with wards, and a lot of zeros in the ED pop file too. Going to need to look at a map to dig into what's happening. Obviously, it couldn't be easy, could it?

Right, yeah, p.4 of the 1991 definitions: EDs below 50 are set to zero, with some complex rules about reassignment. That wasn't the case for Scots OAs in that decade, which is why we were able to use them.

So a small chance this may be quite a bit more work. Given the time, looking at inequalities in Scotland may be preferable. But let's see.

..

Yeah. Let's just read that abstract again... I think I can do it, but I think it's going to require a Java optimisation programme that tests probabilities of where EDs were reassigned, giving them a spread of possible values, etc. It's a bigger job.

It might be, reading the defs, that the re-assigns stay within wards. Hell, let's just have a look at this. First thing I need is merged 91 eng/wales shapefiles.

Then I need to check if the zone names match so I can link...

### Aside: things for meeting today at 2pm

Just a glance over. I was going to send something round but can just go through it. Or send it round.

Where did I get to with the stats part? Well. Some stuff. There's this as well:

"Hur. Just re-reading Geoff's model again: "share of migrants from CoB I" **does** seem to suggest it's **cross-Scots** %, not within-zone %. Which did make sense to me at one point: it's about where migrants are moving/drawn to/pushed away from."

..

So what's the meaning in each case? I think it has to be per-zone %. Otherwise how could the % now depend on the % for other CoB in the previous time period?

So it's probably not cross-Scots. But can check!

..

Err. So let's make a list of some sort. Might as well not try and include owt from presentation. Things to say:

* Presentation on Wednesday - title "Spatial Persistence in Migrant Locations across time: how do inequalities interact? A long term view using UK census data".
  + Work in progress still: will present the method / data and look at how country of birth and employment levels have interacted over time, but only in graphical way.
* Still haven't quite run the model but have all the bits / been playing with different forms of the weights matrix.
  + How to move forward? Visit or just over phone etc? (I'd quite like a visit if poss.)
* Between now and Christmas:
  + Complete Eng/Wales data.
    - The hard part: need to make optimisation program to deduce holes/resassignments in the Eng/Wales 3-census LBS (ward level data). It's trickier than Scotland as EDs also have privacy holes (Scotland's OAs for 1991 don't). Plus side: checks so far with colleagues suggest it hasn't been done, so we'll have something unique.
  + Model running.
  + Paper draft by end of the year with Scots and Eng/Wales data. Can start writing up method stuff in Nov.

### AQMeN / not Eng!

So working with what I've got. What was I looking at before? I'd attempted a Lorenz curve of zone unemployment. Somewhere. Well, it didn't really work anyway - I need a different hook.

So let's look at what we've got and think about it. And please remove any stress and just allow some time to sit with it!

So imagine I wanted to just do this:

* Take the existing boxplots of CoB vs employment in those zones, and more clearly show how each has changed over time, or hasn't.
* It would also be nice to be able to say something about the spatial nature of each of those: did they stay where they were during that? Were they in clusters of any kind? (Which one might potentially work out from a spatial weights matrix to get average amounts in surrounding zones, but...)

OK, so that's good cos it builds on what I've got. But let's also mull how to use it to say something about inequality, which is what we want, right? Well: inequality in terms of employment levels. Though they're also changing...

So. Let's look at it from one angle then another. Non-spatial first, just how employment levels for different CoB groups over time is changing.

How to show that in anything like a clear way? Two things:

* Can just use median / CIs and the three time periods and show groups. (Like windfarms CIs).
  + And break down those groups by most stable / lowest / highest etc. Largest change via SD over time again, perhaps.

So what data am I using for that? R probably gives me boxplot median and quartiles. Let's see.

Oh, another useful thing to do would be: **preserve the same order for the boxplots**. So easily view change over time. And try and keep the y-axis fixed between them.

And rather than doing per TTWA, I should probably start with the whole of Scotland...?

..

Staring at the data trying to work out what to get out of it.

* Can use var **topPercent**. That has e.g. all 3 census years, CoBs, percent employment.
  + It's also already extracted to the top x% of zones by CoB percent (percent across zones I think?)
* X axis to be years.
* Y axis: single line per CoB (picking subsets eventually).

Well that worked first time, rather shockingly. But it needs some more now: I think ranking these different groups to see how they change / how stable etc. There'll be some shifts, but which shift the most? (S'what I did for GRIT innit?)

..

Outputs that I might be able to use from this for the five census rankings:

* Lanarkshire / Glasgow / Aberdeen / Edinburgh.

Might be it. The rest I suspect are too low in their numbers to show anything useful.

..

So what have I got? What do those show, if we're talking about inequality? And! A point about the Scots numbers: those places with the highest number of Scots ... they're just likely to be the least urban, are they? Just trying to figure out why they're always down the bottom of the employment lists...

..

OK, so now I need to pick some out. I could do with:

* Which dropped the most, end to end?
* Which went up the most?

For that to work I think I'm going to need larger zones. i.e. NUTS3 zones. How much of a pain is that likely to be?

Nah, there isn't time. Let's just look at some key places and drop that stuff in. Just need to see for 3 and 5 census change. Possibly.

And write up the presentation accordingly.

### Last bits of presentation

With not a great deal of time given I lost yesterday. Only have something like 3 or 4 hours here. May have some more time on the train or at hotel or as a last resort in the morning. But let's get something into shape here if I can.

Main thing: nail beginning spiel bits so I feel like I know what I'm doing. Though also some graphs for the end. But...

Ah: just slightly fiddly thing. Trying to order other decades by 1991. Different numbers of zones in topPercentYear. Err. OK, done that.

..

So the last thing I need is to pick out some patterns for the five censuses. Which I had. But. Err. Let's look.

Oh, OK, first: let's choose a subset. Most of those places don't work. So. Pick some:

Glasgow / Edinburgh / Aberdeen ... that'll do! OK, now I just need to pick some out, right? Might be possible to put the 3 cities on one slide.

* Aberdeen:
  + India has a steady drop.
  + Pakistan kind of opposite path (increasingly in wealthy areas - has the wealth of the area changed or did they move?)
  + Irish Rep stays pretty stable.
  + Scottish: yeah, big rise. Again, why? We don't know if it's the same people or migration or...
* Edinburgh:
  + Africa/ new Commonwealth: big drop
  + India: big drop in last decade but different to Aberdeen - rise before that
  + Irish Rep **goes up** - again, change. (Might be worth comparing CoB not place)
  + Pakistan: rise, again, but a drop in last decade
  + (poss Scotland if putting on one slide)
  + Welsh never far from top. Something in other cities?
* Glasgow:
  + English and Welsh never far from top rank
  + India: opposite pattern to Aberdeen
  + Irish Rep going up, same as Edinburgh
  + Pakistan doing better than in Edinburgh

..

Well I can probably talk through the ones I've got reasonably well. Just need to change the conclusions a little bit. Might take a bit of time to check on the windfarms one.

### Bits

So got enough for an actual pres / need to do some practicing actual slides. But some time to mull whether to have a little go at finding summat else: inequality / spatial. Something quick, from what I've got?

So let's say: CoB / employment / change over time.

Stick to employment for a moment. Plain ol' maps. Do I already have that somewhere? Well for 5 census, yes.

Perhaps it's not possible at this point: I'd need some measure of clustering, probably can't do that right now.

## Data for Geoff

### What's what

I've sent a bunch once before. It might be that's still mostly correct. Err. Let's see what I said to him! Ah: just for the 5 census. OK. 3 census required. 5 census hasn't changed, I don't think. But. Err.

Let's check what we have for 5 census, see if it all makes sense / remind myself what's there.

The 5-census email. Note the lack of pre-91 house prices.

..

Here's some files for the 5-census country-of-birth analysis. I'll be back in touch in a week or two (probably closer to two!) with the 3-census / extra CoB category data too. Let me know if you need anything else / need things in different formats etc.  
  
So:   
<https://dl.dropboxusercontent.com/u/306562/SMI/Migration/CensusStitching/fiveCensus_to2011_IZs_csvs.zip>  
  
That's three census variables in separate CSVs for each census: country of birth, household count and 'economically active', which contains total 'economically active', total unemployed, and percent 'economically active'. There's no separate total population: I've just been using the sum of CoB per zone to avoid differences to pop counts. (Though accurate population comparison across censuses is a particularly thorny issue.)  
  
For CoB, the country groups got abbreviated, annoyingly. So to translate the unreadable ones:  
  
Rst\_\_UK; Irsh\_Rp; Old\_Cmm; A\_\_N\_C\_; India, Pakistn, Othr\_Er, SE\_A\_N\_, Crb\_N\_C, Nw\_Cmm\_, Rst\_f\_W  
  
That's: "rest of UK"; "Irish Republic"; "Old Commonwealth" (Oz, NZ, Canada); "Africa (new Commonwealth)"  
India and Pakistan are readable  
"Other Europe" (not UK); "South East Asia (new Commonwealth)"; "Caribbean (new Commonwealth)"; "other new Commonwealth" and "rest of world"  
  
This is the house price data, all in one CSV: average price per Intermediate Zone, per census (two year band, one year each side of the census date.) ->  
<https://dl.dropboxusercontent.com/u/306562/SMI/Migration/CensusStitching/1991to2001_twoYearBandMeanPrices_2011_intermediateZones.csv>  
  
This is the shapefile for the base geography used, the Scots 2011 Intermediate geography:  
<https://dl.dropboxusercontent.com/u/306562/SMI/Migration/CensusStitching/Scotland_IntermediateGeography_2011.zip>  
  
This last one contains a couple of possible weights matrices for the 2011 'intermediate zone' geographies, based on different contiguity conditions:  
<https://dl.dropboxusercontent.com/u/306562/SMI/Migration/CensusStitching/weightsMatrices.zip>  
  
Both are row-normalised. One is just queen contiguity. Problem being, islands remain disconnected and Scotland has a lot of these! The other is nearest-4-neighbours contiguity as an example of getting round that - but we can try anything else (like manually connecting islands...). I've been using functions from this, it has plenty of other options to try:  
  
<https://cran.r-project.org/web/packages/spdep/vignettes/nb.pdf>  
  
..

### 3 census

So what have I got for the 3 census / what else needs doing? I must already have EA / house prices / CoB. WHERE? Oh, in really obvious place. The only thing left I need to do: the intersect for dwelling number from smaller 01 and 11 zones. (And find where I put the house price data...)

Lord. I didn't name anything very sensibly. Stare at code time! OK, so it's -->

**3census\_LBS\_dataIntersects.R**

For the correct-counted PCSs. So it's here I should be able to do the dwelling data, in theory, if I can find it.

Where did I do the "aggregate from old to new PCS values"? Err, here I think:

**91\_PSCzeroCountZones\_aggregateTables.R**

Well... I should download the data! I don't seem to have it. Oh yeah I do - just gotta find the right table name for the 2011 bulk download. And there'll be some formatting to do.

Oh hang on: just realised I don't need to be doing the first job - aggregated 1991 PCS data to the new larger PCSs. I need to do an intersect from smaller zones. (I've already **done** 1991 dwelling count!)

So it should be **3census\_LBS\_dataIntersects.R**. Let's look!

..

Aaaah OK, got summat wrong in my 1991 summations. At least I hope I have. Let's look.

I think I was using "occupancy of dwellings and household spaces" LBS65. So why did I think to sum all columns...? Err, maybe cos it's **occupancy**: that'd be why the numbers are so much higher.

Comparing to 2001, it seems to be **just the** **first column in the 1991 table**. I can't really work out why though. It does say it's... well, let's see!

OK, we have a new copy, makes much more sense. Might even be right, you never know. Some of them are close enough to spot on that I trust it.

So I just picked the first column from LBS65, "all types of occupancy". I think the heading is misleading - "total dwellings" selects all but this is like CoB, it's got subtotals within it.

..

Onto 2011: need to find correct table again (and then format appropriately) from the bulk scots download. Let's check KS401SC... looks good. Made copy in folder:

**2011\Scotland\OutputAreaStd\_ALL\KS401SC\_tidy\_justAllDwellings.csv**

So editing down just to total dwellings... OK and all done. And the numbers look plausible enough! I think. I should probably check them against pop numbers really.

..

### Got everything now?

I think Geoff couldn't work with shapefiles, is that right? Need to go back and look again.

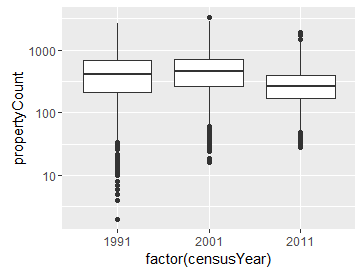
Ah: think the house price data will need re-doing from the latest. I will also need areas for the various zones. Can't remember if we said we're working with 3-census only for now...?

And! I've lost the bit of paper where I wrote down what I was doing. Hmmph. C'mon, this ain't complex!

Well... I need to redo the housing data / check whether or not I did already. Let's look. Where did it get done? **aggregateHousingData\_LBS.R**

Did I use latest data?

OK, spread of sales count in the various zones. Still problematic early on. (There's two sales in that bottom 1991 zone, having skewed the window for averaging.)



### Country of birth oddness

* Two fields: "Other middle east" and "other middle east" again.
* In 1991, different values
* The other two, same values (prob due to re-assigns)

So what happened? Well - I don't know. But I've just re-run and it's all OK. Phew. Fiddly! Err. So what now?

I've got CSVs outputted for everything? Err yeah.

## Extraz

...

### Probability shizzle for CoB vs employment zones

Drawbacks being don't know the ecological wotsits but let's work out the basic version (Which I just failed to think through for wind turbines but hey, let's try this).

So this is a way to get a spread of results for employment levels per CoB without picking on deciles. Plan:

* Zone x has 85% employment. 10% Indian. Pop 1000 so 100. (Actually see below, just assume this is EA number.)
  + Oh argh: also have to assume equal 'economically active', which won't be true either. This could be a bit silly. Or good practice prior to better microsim? Mebbe.
  + So we'd need EA% for whole zone then some assumption about how that maps to each CoB.
* Assuming we don't know who in those 1000 are actually employed, we can do random picks. Repeat to get spread.
  + R prob has an easy way to map probability to proportion; this is gonna be some max likelihood typa thing.

# Pulling things together

First proper look after the meeting. Been through notes once: going to take me a little while to get back into the zone with this. There's a lot to think about. But some things to start with:

* Get code to the point where it can output Geoff's final data form. That isn't too much effort and should be useful all round.
  + (Though I still want to be doing these regressions myself.)

Oh and something else, from my previous explorings of the regression:

* I still think it's possibly odd correlating zone% between time periods. If they stay the same, perfect correlation. If they shift...?
* Which then brings up the point: shouldn't you be comparing shifts in different groups? i.e. in places where the % is low, has it dropped or increased? Same for high?
* Now I think about it, that's not a bad approach overall - a focus on change. You could link it with the weights matrix. So e.g.
  + Find change per zone between the time periods.
  + Hmm ... will think about this in a moment. There are a lot of options. Just mulling if % stays same, how you'd compare to what's happening nearby.

### Writing to Liverpool

Remind myself what I wanted to say and who I'm writing to.

* Chris Lloyd, Gemma Catney
* Planned RSS seminar in London... March/April.
* Interpolating the zero zones.

### Tracking down Scottish economic history

Will need some stuff on housing demolitions too. But for now:

<https://en.wikipedia.org/wiki/Economic_history_of_Scotland#Deindustrialization>

See those two sources down the bottom. Library seems to have both...

Yup! Downloaded one. Here's the article:

<http://www.euppublishing.com/doi/pdfplus/10.3366/sesh.1995.15.15.66>

..

Some links:

<http://www.scotsman.com/news/steel-industry-in-scotland-a-short-history-1-4243119>

Note **nationalisation** there. Plant nationalised in 67, privatised in 88.

..

Dalzell: closed by Tata, re-opened recently. Scotland's last steel production.

<http://www.scotsman.com/news/sturgeon-to-hail-future-of-steel-as-dalzell-reopens-1-4242474>

This stuff could be covered by the BSD.

..

This is looking pretty thorough:

<http://bellacaledonia.org.uk/2015/11/08/scottish-steel-a-lesson-from-history/>

So this is really good on: links between coal/iron/steel / types of steel production / ship-building / later car building once stripping milling was in places (govt funded I think). Then you get Hillman (Rootes is the name of the company; Chrysler later bought them out - 'modern scotland' p.327) in Linwood, opened 1963. and British Motor Corporation in Bathgate, opened 1961.

"Linwood closed in 1981, Bathgate in 1986 and Ravenscraig in 1992... These and other closures have been called Scotland’s industrial clearances."

So that's enough for some local shocks between those two censuses (and another... big gaps, mind.)

"To follow the Brassert plan and build a modern integrated coastal plant producing steel for shipbuilding, steam locomotive building and other heavy industries would have risked turning Motherwell, Coatbridge and Airdrie into ghost-towns."

### Finlay book

*Finlay, Richard. Modern Scotland 1914-2000. New edition edition. London: Profile Books, 2005.*

Chapters here, downloaded via library link.

C:\Users\SMI2\Dropbox\SheffieldMethodsInstitute\Migration\Readings

..

"... manufacturing’s share of employment in the United Kingdom (34.8 per cent) and Scotland (32.4 per cent) was a good ten percentage points behind that in West Germany, which was cast as the example of the successful economy." [323]

* Bit of a surprise that UK higher overall.

..

"The period was also marked by a number of highly publicised closures, such as that of Plessey in Glasgow in 1971, though most jobs were lost as a result of cutbacks and streamlining. Scottish unemployment figures crept upward from 80,000 in 1970 to 130,000 in 1972. Although there was a sharp fall thereafter – back down to just over 80,000 in 1974 – the figures then climbed again just as steeply, peaking at over 160,000 in 1977, where they stabilised until the end of the decade, only to rise sharply again in the early eighties." [327]

### Looking into the data

An obvious thing to do from this: **look at changes in the mapped employment / pop data** to see if there are any obvious changes between decades and where those changes are more severe. Though... there's a tricky spatial stats issue. Mull. Change compared to mean change etc.

..

Given the amount of comparing that needs doing, it's going to be easier to do in R. Though then actually mapping it might not be. So what am I looking for here?

* Areas where employment changed more than background levels... looking for losses of industry. Picking out how patterns are different.

Should be able to map those patterns, right?

<https://www.r-bloggers.com/spatial-correlograms-in-r-a-mini-overview/>

back to spdep? Err. Well, I'm totally not going to be nailing this in one sitting!

<https://cran.r-project.org/web/packages/spdep/vignettes/CO69.pdf>

..

Let's pick some questions I'd like to know answers to. Well, first...

OK, can I now see change between censuses plz? Hmm, nice boxplot.

Oh - I'm working in:

**censusDataLookz.R, section five census econ active**

And now I can see the differences in change between censuses, I'd like to know if drops are happening in the same areas. I guess I can probably look at that in one of them map things.

..

Well. I think after 71-81, a lot of the changes are likely to include a lot of demographics n stuff too. I doesn't look terribly easy to say much about shocks. So that's a great conclusion.

### Code for Geoff end output

Just to give me something to do to get back into this! So let's look at Geoff's files. Stuck it in:

C:/Data/Census/Geoff/countryOfBirth-gm.xlsx

..

OK, so it'll take some piecing together. Things I know:

* The final has each CoB long. With no direct indication of which is which, but I know it's been reduced to 22 categories. Which I think...
* Oh yeah, it's in **2011-stacked**.

So what are all the fields? Should be able to deduce that fairly straightforwardly. Oh, and some basic diagnostics from any batch (like wot Geoff wuz doing). Let's have a look.

So the final thing's got no gaps between CoBs (we know which ones from previous sheet so can hopefully deduce how those were summed - can use existing function to recombine, right?)

There's one var that's confusing me. Can I rule the others out and deduce?

* **Xij1991/2001: migrant shares**
  + Note: should set up so can easily create for different... yeah, need to reorganise the stitching.
  + Doing a base stitching seems sensible. Later aggregation can be the same for all decades.
* **Migsh91**: "migrants from other countries" - though currently just "all migrants"; own group not yet subtracted. (I can do that.)
  + So presumably minus Scots but not English? Or?
* **Ph91:** house prices.
* **Pop91:** err. Population.
* **Hs91:** dwelling count.
* **Pophs91:** crowding / people per dwelling
* Then log house price and log crowding.

..

Oh hang on: it would appear that it is, in fact, doing shares **across zones, not within zones**. Yeah, we've got 100% summing going on across zones in the 'migrants reduced' sheet.

Ah. Now I read the paper, this makes more sense of the regression. The first line p.252 in the print-out:

* "The proportion of migrants arriving from county (i) who locate in any local area (j) is a function of the existing stock of migrants from (i), the stock of migrants from the same country of birth in contiguous areas, the stock of migrants in the area from other countries and a set of variables..."

..

OK, so that makes it easy enough to set up, in principle. I'm still mulling what it means, in a usual overthinking way. Somewhere above, I've got some discussion of the value of looking at % per zone vs % across zone.

But the point of the above:

* Assuming some fixed proportion **will** appear in the country, the regression is then working out what their location choices are correlated to. (Which is, note, doing away with the question of whether the **number** of migrants is endogenous. It's not talking about that.)

..

Ah: the problem was:

* **Larger proportions are going to end up in the larger population areas anyway, aren't they?** That might not matter so much for London boroughs (perhaps?)

Let's just check all that in R huh?

..

### More on getting Geoff-ready output (and some other things)

So:

* Mystified by the use of shares across zones. I can sorta-see why but... population! Surely?
* Anyway: it means I now, in theory, know how to knock that final sheet together. There are still a bzillion prior stages and I haven't made those tidy or even comprehensible yet, and really I could do with doing that for my own sanity.

..

Back on this. Silly block. What I'll do:

* New R script plz, **eViewsReady.R**.

Now, what data did I need? Um. Might as well reload everything saved. Which was what? (And note, the employment stuff got one backwards. Can I correct that before trying to recombine?)

..

OK, fixed econActive. All a lot of data processing - going to be more when it comes to England, huh?

Now to reload all that shizzle and reshape into eViewsNess. I think I also want to just run the regressions in that file. But let's load first then see.

Nearly there. Last thing: what's migsh91? Didn't I already work this out?

Nearly, and I can help Geoff out here. So what it's doing:

* Migrants from other countries: their share of each zone. I think.
* Why do I get the feeling it's just the sums of the proportions?

Newp, can't find what was done. Oh, yes I can. Duh. So now it's all non-UK born (not just non-Scots, which I'm not sure about.)

..

Before that, let's fix CoB and make sure the double counting's right.

..

### How 91PCS process worked

Having to remember all this. It was a faff.

* Stitch CoB categories together (column sums)
* Process 01 and 11 through intersect
* 91, process separately by aggregating to appropriate zones (in its own R script)

Leaves you with 3 shapefiles in:

**C:\Data\Census\StitchOutputs\Scotland\LBS\_postcodeSector\_3Census\_raw**

These I can convert to CSVs. Where did that happen?

..

OK, so I've got back to where I was, having remembered, almost, what everything does. Lord, it could do with being organised. Effing mess.

Now for the original migrant share numbers. Then the ones with the own-group taken out.

## Running regressions on that data

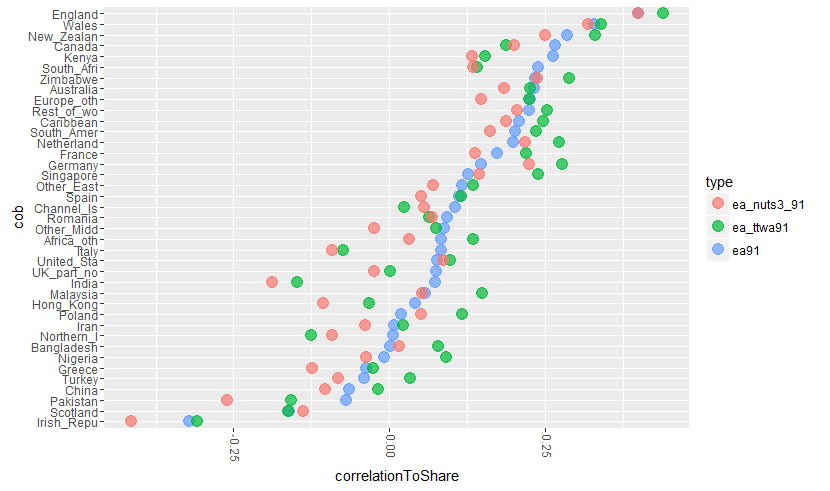
To see what's what.

### Larger employment aggregation

A little bit of gentle xmas work, I'm just going to aggregate employment to TTWAs for a more sensible employment regression. It might actually be easier just to do centroids – then I can do in R and add NUTS3 too.

..

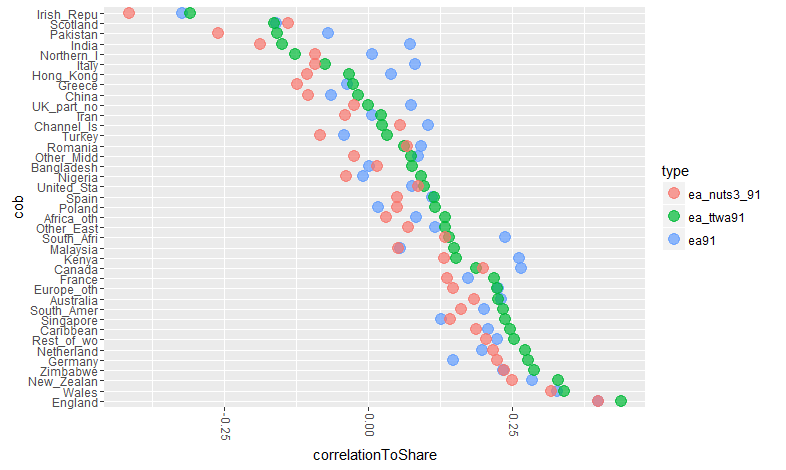
OK:



So to explain: y axis are r-squared values for each CoB against employment % in that zone – for direct zone level, then also for each of the aggregated zones. Ordered by direct zone; might try the other two in a minute to see if anything else jumps out.

Some of the patterns we've seen before.

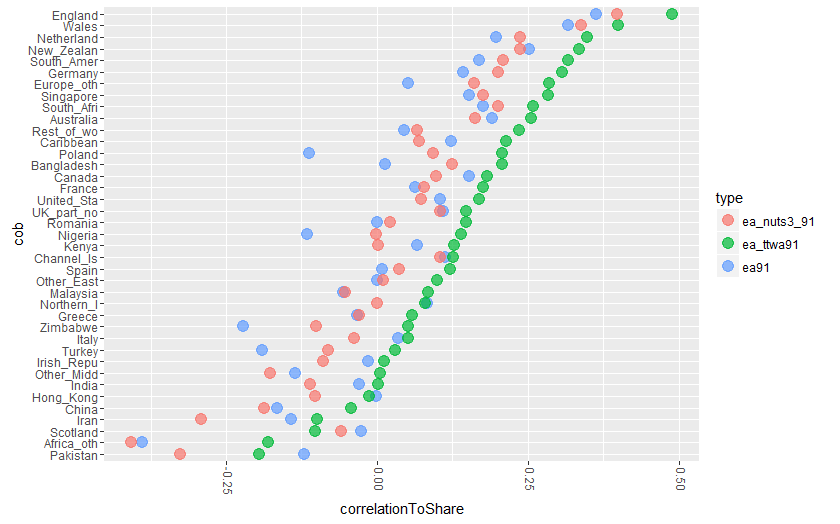
Here's by TTWA order (coords flipped too):



So it's looking a lot like a plain proxy for wealth again to me.

..

Oops: **those were all 91-cob correlated to 91 econ-active.** This is 2011-cob shares to 91 econ active (ordered other way round too):

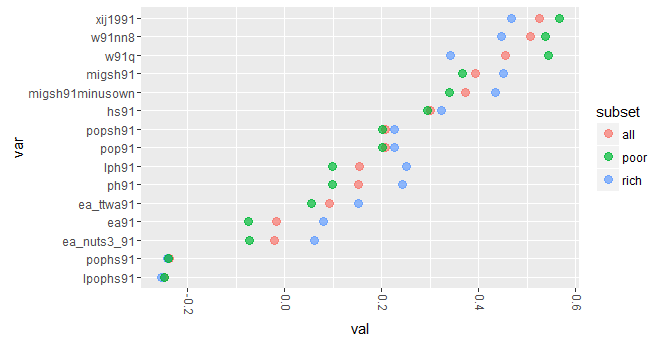


So there's a **much** more clear pattern of TTWAs doing a good job of capturing location choice / employment interaction. Out of interest: what happens with 2001 CoB...? Nuffin very exciting.

...

Done rich/poor. Regressions OK. Plotting the patterns it doesn't seem very clear – needs breaking down by CoB I would have thought, then combining.

Then – the following: this is all the r-squared values for all, rich and poor. For all vars. So it starts to give a useful overview. Next: some more breakdowns by CoB – e.g. I'd rather see the mean r-squared for "rich" for each of the CoBs within the rich group, not the whole lot. But let's check the plots.



### Sending something

So let's get all this together so I can send something round. Now I'm back in the office I can check if the regressions are looking roughly correct. (They'll be an eensy bit different.)

..

Yeah, the regressions are looking fine. I've clearly not quite got the correct 10 rich countries but the others look OK. Actually saving in a nice format?

<https://stat.ethz.ch/pipermail/r-help/2007-November/146799.html>

..

OK, I've got a bunch of regressions output in a reasonably readable output. Now just to remind myself what I'd been looking at with the above outputs. A couple of things...

## England 1991 PCS

Because why not? It's a fun thing. Just now that seems important. Need to try and remember.

Reading the previous section... then the definitions. According to the definitions, the ED values **should somewhere** have "a count of persons" in full, even for those where SAS values have been re-assigned. Hum. Best have a look at this.

..

We're in: **1991zeroCountWorkingOut.R**

And also in QGIS / engwales1991\_geogcomparison. Where I'm intersecting to get towards an ED/Ward lookup. (Don't currently have.) It's going to take a while to run.

Though I'm not sure the rationale will work in this case if we don't know for certain what the ED values are.

So I need to remind myself of the rationale:

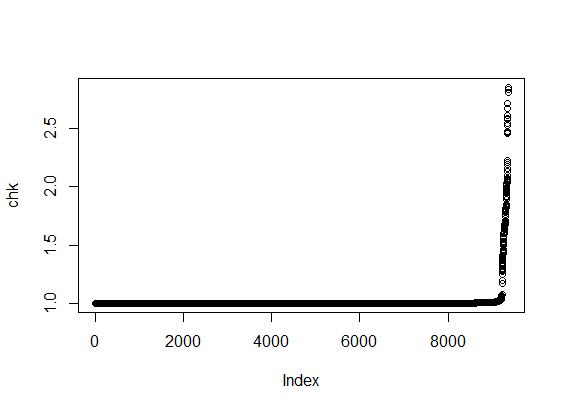
* Aggregate ED counts to wards (in this case, assuming that most are going to be in the right ward, even if ED values have been re-assigned - they should be in other EDS still in that ward.)
* Put those counts against what we have for the ward count.

Oh silly me: **the ward code is nested in the ED code**. I knew that before, right?

..

It occurs to me now I know a bit more about how to make them work, there must be a way to use a contiguity matrix to make this easier. Though... that's a rather large matrix innit? Well, 660mb.

OK, so this: ratio of LBS ward popcount and aggregated SAS EDs. No perfect cutoff. So somewhere in that right hand side, our candidate differences: larger ward pops than we should expect, so where numbers have been moved.



So we want neighbours that, when re-assigned, are \*closest\* to this. Problems:

* The numbers are not precise, as they were for Scots OAs/PCSs.
* Reassignments often happen from e.g. two contiguous zones to one.

So how best to look at this? Well, need to keep on reminding myself of the rationale, so let's go again:

* If LBS wards < 1000, their count is assigned to a neighbour.
* More than one <1000-zone may be assigned to a single neighbour.
* ED popcount summed to ward-level: if it's **more than** the LBS ward, this is an indication **this ward received a count from a neighbour.**
  + Hence we look at where LBS-wardCount / EDpopCount > 1: the above plot.

So we're:

* Picking out LBS wards to fit together
* Cos LBS is zero but ED-summed is not
* And we have (approximately) that ED value in a neighbouring zone if you subtract the one from the LBS zone.

Ah, which starts to offer an approach to using the contiguity matrix, possibly. **Let's stick this shizzle in a map and see.** Need to mark some things up. With these details:

* **Zones with zero in both** are in fact zero pop, I am assuming. E.g. city of London, other commercial.

Dammit: just realised I'm only currently using England data. Will need to join Wales up too.

Update: newp, I have! All good. Got confused by shipping.

..

Ah: **we can also use contiguity to work out which groups of LBS zeroes are likely to be combined**. Possibly. Will need to look at that.

OK, map is helping. Map is, by the way:

**Census/QGIS/engwales\_1991geogcomparison.qgs**

..

So looking at it, I think it's worth doing what I did before, but trying to use the contiguity matrix (because the values don't appear to be so simply in the next slot.) That should set things up to then try other groupings.

I also need to work out exactly why there's an ED / Ward shapefile zone count difference. Shipping is a lot but there's something else. (I mean the EDs aggregated to wards via 6 char part of zone code).

The ward **data** has the correct number in. So. Wassup?

..

OK, done some number checking and dissolving of the ward shapefile to single IDs. All good once shipping removed. (There isn't even a nominal shipping zone in there.)

So now back to the task at hand.

### Actually finding the reassigns

I think I can use a contiguity matrix. And probably do in stages, starting with single pairings. I also need to account for / examine near-pairings - the values may not be exact.

So a couple of possible approaches here. Note, we're starting with the known-zero-set wards and getting its neighbours. I might not even need a matrix - the neighbour object should just give me the list, shouldn't it? Since the job is to identify...

Well, let's play.

..

OK, that's easy. Got poly2nb object with indices for each neighbour.

..

Code for matching on pairs only has done pretty well. Now (after a quick check to make sure it worked) to look closer at the remainder. Which are multiple re-assigns? Are there any with non-exact numbers?

For multiple reassigns, we'll need neighbours of neighbours - but that will just be combos of two or more neighbouring zero-zones.

Yup, current looking good. Now what about some others?

..

Central London on the Thames: I think that area's going to be tricky. There's something different going on to everywhere else. Let's look at other places first then come back to it.

Other odd things in London: zero-count via ED agg, but with positive LBS. Hmmph.

..

And! Mostly it's two-contiguous-zero-zones into one neighbouring zones. (Or I've found one example.) But there is one case of a single-pair nearly match. But we should probably do those last.

So first-up, work through two / three / more contiguous zero-zones and look for the sum of their missings. I don't think there are many of these.

And have to think: I'm not sure they will always be contiguous to each other. They both would need to be contiguous to the zone their counts got added to, so that's one approach I think.

* Note that I'll need to not double up: if two zero zones are going into one other, don't find values for both. (Would need some way of removing.)

# Starting on the paper

## What's actually going in the paper / what can I write right now?

Just to get things going. So rather than getting stuck into (stuck on) the lit stuff, there's plenty else to be doing. Let's get that done / planned out. Let's even start a document for that. Crazy. Word I suppose. (\*sigh\*).

OK, so I've made a blank document, and the blankness is staring at me with a disturbing look in its eye. How about a bit of a breakdown of the tech things that need writing up. I shall call it ->

### A bit of a breakdown of the tech things that need writing up

* Harmonisation of CoB categories. Explanation of the differences, 3 vs 5 census, choices made there and why.
* Harmonisation of geographies.
  + Why different for 3 vs 5 CoB.
  + Creation of larger PCS to make zero counts go away (quote the census defs to show contiguity)
    - So say something about the average stats for those different zones.
  + Citing other approaches / Walford etc.

I had to re-do this recently, didn't I? Which way round did it have to happen? Oh: CoB harmonising first - so aggregating columns in their original zones, before then splitting those counts according to the re-assign values.

How to write up in a way that makes it clear? The things to get across: **why** 5 vs 3 census CoB is different, why the different geographies.

Hum: I think the two scenarios/datasets need their own names to make sure it's as super-clear as possible. Err. What's good? The 5-census and 3-census datasets? It'd be nice to have something slightly more descriptive. I should probably attempt to write something.

## Scanning related gubbins

### Copied over from urban indicators doc (Googling "impact of migration on employment and housing")

<https://www.theguardian.com/housing-network/2016/jan/25/is-immigration-causing-the-uk-housing-crisis>

Some links via there. Also...

<http://www.migrationobservatory.ox.ac.uk/resources/briefings/migrants-and-housing-in-the-uk-experiences-and-impacts/>

* Last point important: "Research on the impacts of migration on housing costs has not reached a firm conclusion."
* "Costs of housing take different forms – notably house prices, the private rental market and (low-cost) social housing – affecting different groups in different ways, depending on the type of housing they are most likely to occupy."
  + So using straight house prices / interaction will be tricksy.
* References this: Belfield, C., D. Chandler and R. Joyce, ‘Housing: trends in prices, costs and tenure.’ Institute for Fiscal Studies, Briefing Note BN161, London, 2015.
* "Evidence on the impacts of migration on house prices in the UK remains inconclusive. Some studies have addressed this question by comparing house prices in areas with lower and higher levels of migration. They found, counterintuitively, that migration to a local area led to a decrease in house prices. Using data from 2003 to 2010 for England and Wales, Sá (2014) found that a 1% increase in the stock of the foreign-born relative to the local population led to a decrease of 1.7% in house prices. A similar result emerged from a study by Braakmann (2013), who also found that price decreases took place primarily at the bottom of the distribution (i.e. in less expensive housing)."
* "However, these results are explained in part by the out-migration of UK-born people from areas with increasing in-migration. The Migration Advisory Committee (2014) has argued that while migration may be associated with house price decreases at the local level, out-migration of UK nationals to other areas could mean that migration increases house prices, on average, across the UK as a whole."
  + Right, so: what I want to know: is that just normal churn / change in high throughput areas? What defines high throughput vs... is this related to migration efficiency? Will have to remind myself what that is.

### CEP / Brexit + impact of immigration

<http://cep.lse.ac.uk/pubs/download/brexit05.pdf>

"The best source of data to explore the impact of immigration is the Labour Force Survey (LFS). This is because it makes it possible to measure the economic circumstances of people born in the UK and compare them with immigrants from the EU and other countries."

Oh, other bits of this are in the Urban Indicators doc int they?

Ahhhh:

"There is a huge amount of research examining the effect of immigration on jobs and wages (summarised in Wadsworth, 2015; Portes, 2016a; Centre for European Reform, 2016; Dustmann et al, 2005, among others). The conclusion of this research is that the large increase in immigration in the UK has not significantly harmed the job and wage prospects of UK-born workers."

***<end of urbanindicators copy-over>***

### Random other-paper thought

Something else to get out of this that just occurred to me: a paper comparing the worldview differences between quant migration and dealing with it on the ground. I'm thinking of how radically different Charles Clarke's take on what works is.

So not just a critical paper but asking: how can these two quite different ways of seeing be supportive of each other? What are the problems with that?

Getting ahead of myself, but that would be interesting for a year's time.

### Scanning the scanning

So this is going through 'migration modelling scanning' with an eye on starting to write this up. There's also a bunch of new stuff I'm surveying in the urban indicators doc. It might be better off here, not sure.

..

Quick read-through first just to get some ideas zinging about the place.

Good timing:

<http://blog.policy.manchester.ac.uk/posts/2017/01/segregation-and-inequalities-what-should-we-take-from-the-casey-review/>

Loads of great stuff to bounce off. We definitely haven't really tackled the issues of using CoB compared to this stuff. Probably I can get some OD flows for CoBs just between the whole of Scotland and the other nations, right?

Cf.

<http://www.socialintegrationappg.org.uk/>

..

**Back to the doc.**

<http://webarchive.nationalarchives.gov.uk/20160105160709/http://www.ons.gov.uk/ons/guide-method/method-quality/specific/labour-market/labour-market-statistics/index.html>

Just trying to track down LFS questions. Kinda forgotten LFS/APS relationship (the latter supersedes the former, right?)

It does look like there's a national identity question.

Oh: and for flow data:

* **We would only need to break flow data down by country to know who'd come into Scotland from rest of UK**
* And could then perhaps break down Scotland into more detail

On the point about separating "CoB from abroad" and "CoB already here but hasn't moved or only moved internally": given we're looking at how they **cluster**, does that help? That is, can you say something about migration patterns by making some assumptions about that clustering? Ie.:

* Assume x% internal for this CoB, y% external ... any reason to expect different clustering patterns for the two? If not - or to the extent that they're similar - is there any problem at all in **not knowing** whether CoB count was already there or is new?
  + See point above about flow data: might be a simple way to narrow it down a little.
  + Oh: we also don’t know how many died or moved out of the country presumably. Hmm. Need to note down all of these. Like I'm doing now.

Or ask a different question:

* If we **assume no new people from other CoBs have entered**, what would any change in clustering look like if we only examine share change for remaining internals?

OK, so we're **not actually** looking at migration flows. It makes sense for **flows** that you might want to use larger zones (something like TTWAs possibly ideal) or the Family Health Service Areas Fotheringham et al use.

But we're using small zones because our questions are about **spatial stability and spatial dependence**.

..

**Random thought**

We're just modelling how / to what extent CoB numbers (shares actually, stocks of) depend on a bunch of factors. Whereas cf. Fotheringham et al are all about out-migration. So rates of movement between healthzones.

**Moar scanning**

So reading more of Fotheringham: determinants of migration itself seems unnecessary for this paper. Here we're only talking about whether migrants from particular CoBs tend to stay in the same areas / what the spatial patterns are / how they differ across groups.

Keeping the focus on that allows later saying "we know much of this is underdetermined (CoB from abroad or already there? Well we can count can't we!) but we can still say this about the **difference** in patterns."

That may of course be due to differences in how existing CoB folk move **within** the UK. (Overlapping with age / demographic diffs - including e.g. older CoB groups now showing older-people migration patterns vs new young incoming Europeans having a young-person migration pattern.)

..

So it might be an idea to use actual total numbers. Particularly if we can get CoB migration numbers for the nations for some decades. I might just check that... uuuurgh maybe not right now. WICID is hideous.

..

**Location quotient**

Just been re-reading:

Stenning, Alison, Tony Champion, Cheryl Conway, Mike Coombes, Stuart Dawley, Liz Dixon, Simon Raybould, and Ranald Richardson. ‘Assessing the Local and Regional Impacts of International Migration’, 2006.

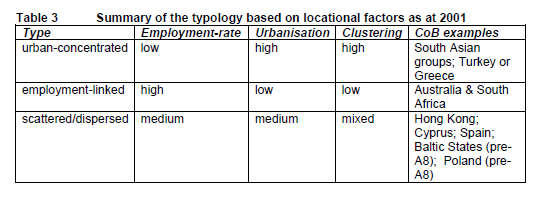
..

Location quotient, pp.20: "What the LQ shows is the factor by which an area’s share of people in that particular CoB differs from the area’s share of England’s total population, indicating where the group has a greater, and where a lesser, presence than would be expected."

Yeah. Makes sense. Either side of 1. Also gives a way to compare change between censuses perhaps - perhaps! Yes, that works, doesn't it? Works a little like shares but can say something more about how things have changed.

..

Note also its go at clustering: not actually spatial clustering as such. pp.26 - typology by CoB group. Could nab that for checks on correlations...



**More on Ludi Simpson et al / Pupil database project ("Measuring Internal and International Migration from the National Pupil Database")**

Within England: finding the most immigrants to be African...? In London anyhoo. Scotland not the same then huh?

Picks up on "transnational family formation" - South Asian. Also poorest? Related or not? Different reasons for moving in the first place I suppose.

"Among the smaller cities and other urban and rural areas, the origin of most pupils likely to be immigrants is more often ‘Other White’ or Eastern European, following the first accession of Eastern European states to the EU in 2004. (Chapter 11)" [6]

"The geography of immigration is changing overall and for each ethnic group. Immigrant pupils are a little less concentrated in a few districts than are existing pupils of the same ethnicity. For most origins, settlement in Britain is associated with a growing spread of residential areas." [6-7]

Interesting that ~13% of those likely to be immigrants are gone again in a year. Cf. question of how much CoB count is reduced by emigration.

## 5 census with PCS common geography

Just getting that done. If RStudio will start. First thing, actually: **downloading the 1991 PCS-level SAS data**. (If Firefox will start.) And may require some processing of the 2011 Scots data, if I didn't already do that at OA level. I think I did. Will need to check!

OK, 91 SAS at PCS level acquired. It needs processing / reducing / giving CoB names.

Processing: a few stages. Let's see if I can find the matching LBS code. Oh good, messy!

* **CoBColumnNameUpdate.R** seems to be the place.

Done that. Now got:

**1991\_Scotland\_SAS\_CoB\_postcodeSectors\_noMaleFemale\_countyNamesAdded.csv**

Next: reduce that to the correct CoB column number **and** aggregate the PCS zones. Can't remember if there was a necessary order - don’t think so. But let's looksee.

Urgh: shortly, there's a whole bunch of stuff in miscDataReshaping that's not been done for 5 census. Hopefully not too much of a ballache. We shall see!

**..**

**CoB\_stitchingCategories.R**

Which is doing the 5 census. Some things:

* Pre-PCS aggregate, yes.
* There are two 2011 versions: one of them's the separate OAs, so can use that. I think.
  + Not doing both though. Which is fine, it won't overwrite the other.
* SAS PCS is the same layout of CoBs as the other SAS, so should just be able to use that. Or even just replace the other SAS? Hmm... Oh, except I do want to keep the previous lot to compare to, probably. So maybe not.
* None of this will make sense in 6 months' time...
* So I'm sticking on the end, not quite in order...

Seems to be working. That just leaves getting these into common geographies. And getting the 91 SAS PCS aggregated correctly. Wherezat? Err. Probably need to nab a bit of it from the 3 census shizzle don't I?

OK, so this was the original attempt at 5-census PCS (when still 1002 / 1003 or so with shipping):

**censusData\_smallToLargeGeogs.R**

It's now:

**censusData\_smallToLargeGeogs\_5\_censusTo1991aggPCS.R**

And I'm going to change it to the agg-PCS zones. 1991 will need the separate aggregation treatment from the 3-census small-to-large which, I think, has a completely different name.

..

Oh shitting bollocking fucknuts: **we have to do it for all variables.** Forgot about that. Many many stages to go through yet! What a lark. Anyway, best press on with CoB to start with before working out what else needs to happen.

Oh wait: **I already have all that for 91 to 11.** Those are the same. Yay! Just need 71 / 81 to be PCS'd. (And to figure the issue with dwellings, which I think probably can't be fixed. But we can do other things.)

So that's nice, innit? Def a good idea to be doing this even though it's a PITA.

..

I think the 5 census other-data stuff got done here but could be wrong / missing some things:

**household\_n\_employmentCensusDataTidy.R**

Will need to check exactly what I'm aiming for and make sure I've got it all. Which is just what I've already got in the 3 census data. Aaanyway - CoB first, yaaah? One step at a time, yaaah?

I will need one new intersect: 2011 OAs over the PCSs. Might already have it but don't think so... Oh no, of course: they all need redoing cos it's different PCS scale. Orrrr... perhaps not. I already **have** 01 and 11 intersected don't I? Yup yup. Just need 71 and 81.

..

It's at this point I need to get the 91 data into the 822 zones. Where did that happen? OK, so that was:

**91\_PCSzeroCountZones\_aggregateTables.R**

Where I've actually written the shapefiles to the shapefile raw folders. I'm amazed I'm managing to track my way through this.

..

I'm yet to run the other (four) intersects. Brain slightly dead. Let's go do the thing then come back and see before running. Might leave the building...

Back! Just run what should be all the CoB raw shapefiles for the 5-census/PCS lot. Waaat next?

..

### Tracking down the remaining variables / working out what next

Remind me what I actually need. First up:

* Nothing else for 1991/2001/2011. I already have it from the 3-census stuff.
* Econ active for 71 / 81 recast to 91 aggPCS
* Really don't seem to have dwelling count so that might be all we've got. What did I actually stick in my final regressions?

..

Note: **we still don't have density** **in any of the data** (I don't think) as I was gonna look into some way of finding out whether the border of housing areas were.

Just put 91-11 dwellings/employment in the **raw** folder. At some point it occurs to me it'll be easier to put all of them in one folder: 3 and 5 census. But for now keeping separate? Guess so.

Or maybe not. Setting up the regressions is going to be easier if I'm only referencing one folder, not several.

Anyway, next jobs:

* Employment 71/81 re-assigned to 91-aggPCS.
* Check on dwellings again for those two decades (don't think we have them).

So let's do employment in the same place I was doing everything else. Which was in?

So it's currently effing messy. When there's time (which may be never) it needs a bunch of refactoring. So let's try and just note where things are and where I'm doing the next bit:

* 3 census: small-to-large geog - CoB, employment and dwellings being reassigned in the same doc:
  + **censusData\_smallToLargeGeogs\_3\_censusTo1991aggPCS.R**
* For the **5 census**, employment n ting were done elsewhere originally (when IZ-related):
  + **household\_n\_employmentCensusDataTidy.R**

So can I use that code? Let's looksee. Yes I can! OK. Now I just want to have another little look at..

### Dwellings / households / difference

Let's have a look here:

<https://www.nrscotland.gov.uk/statistics-and-data/statistics/statistics-by-theme/housholds/household-estimates>

Links to this ("gives information on the methods used to produce estimates of households and dwellings in the four UK countries"):

<https://www.nrscotland.gov.uk/files//statistics/household-estimates/household-estimates-across-uk.pdf>

..

OK, so that's as unhelpful as one might imagine. So this is annoying as I could really do with understanding household / dwelling stuff. But it's probably going to be a massive PITA. Hmmph.

..

71 defs: "household space is used to describe the totality of all rooms used by the household for living purposes..." Oh God, that goes on to be possibly the most insane sentence in the known universe.

OK, here's my conclusion:

* An argument exists for using households to work out overcrowding, not dwellings (cos many/some of those might be empty).
* But I haven't really got time to work that out right now.
* So let's press on with what we have, get the rest done.

### Next steps

Getting from raw to CSVs to eviews sheets (which I can use in the regressions) to regressions. So various other bits needed include the spatial weights matrices for ... whatever reference decade we're using. Which is a good point - what **are** we using? I've currently only made some for 1991. (Good lord this is getting messy huh?)

Well let's make employment CSVs for 71/81...

You know, it might be an idea to do that thing I was saying about just putting them in one folder to deconfuse this shizzle just a little. It'll mean folder output names are wrong for what I've just done but I'll cope.

Yeah. Really - if I can just do one slightly amended version, I don't need to be duplicating code. I might worry about households later.

..

Well. Let's just use that one folder as it is, keep the name the same, maybe hide the others later:

**LBS\_postcodeSector\_5Census\_raw**

Except we don't have 5 censuses worth of e.g. dwelling data. Sooo.

Note: I am half-way through re-writing PCS 91 CoB so the format matches the others for rbinding. Lost my place a bit. Back to it shortly.

Actually, it's probably going to be easier to do that with the CoB csvs themselves.

..

Right: COB and econActive done for 5 census. No sanity checks yet. Should probably do some sanity checks. What else? Weeell I can assemble the basics for the regressions.

### Note about writing

Note for writing tomorrow:

This is about international migration and its change in pattern over time in the UK (Scotland). So that's obvious, right? Distinguish it from ethnicity studies and / or segregation for now.

So note other sources: passenger survey etc. How census CoB is the only data with good spatial resolution and ubiquitous cover (LFS too small in number to get high resolution - though maybe?)

And then some simple questions about that. But this is an obvious point for write-up I've missed - the focus on intl migration plus geography plus over time. Plus why difficult because categories.

## Bits

### Playing with delta employment

In **regressionOnEviewsReady.R**: cos Fotheringham found little correlation to actual employment levels but high to employment **change**. But that was correlation of actual flows - not what we're looking at.

And there's some interesting patterns if looking at the actual data but it'll be tricky to pull out and won't be much use in actual regression lines, wouldn't have thought.

Though note: the dynamic of richer group tending to have their shares places where employment **didn't** change - I think just means those were the wealthier areas with more jobs to start with. Kinda makes sense.

### CoB domestic vs incoming

This is easy, I realise:

* Get CoB counts for GB as a whole
* Get CoB counts for Scotland

Aaah the only problem: we don't know who **left**. But it'd be a good sanity check to see if numbers increase. Yes, essential the more I think about it.

..

### "How to talk about immigration"

<https://www.theguardian.com/politics/2017/jan/19/crisis-of-statistics-big-data-democracy>

So the stats completely not trusted! OK then. So probably not worth us very much thinking about how this work can be used to persuade.

### Restate assumptions

I could do with having a little step back and think about what I'm doing. Haven't really written anything - I could have a go at doing that in a bit. I would also like to have a bit more of a think about the numbers themselves and what they mean.

I was just thinking then, for example, about shares:

* If you have x vector of shares for 1991, and then (say) add the same number of people to each zone - the shares change, don't they? Even though the actual shares **coming in** have been spread evenly - if those CoB folk from the previous time period **haven't moved**, then the share numbers will have altered.

That kind of thing. Need to work out the details of what the numbers mean and what regressing them means.

The other option, apart from shares or per-zone-proportion, is the raw count, of course. But note again: only per-zone-% avoids auto-correlation with population count in those zones. I think.

On the other hand: I have to **try** and avoid the tendency to fall down a learn-everything-understand-everything hole.

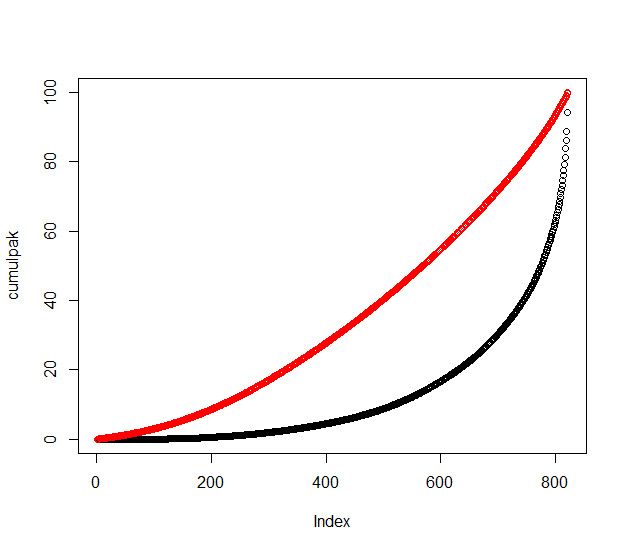
OK, so that hasn't presently worked. Other things:

* 5 census data done. Possibly get household count too (as approach to overcrowding.)

Hum - maybe just get one little thing done tonight. But what to pick on? I think I'd like to think through the numbers a bit. Shares vs other typa stuff. Err. Sooo how do I do that?

### Digging into shares / CoB numbers

So far, so confusing. (Brain not on top form.) But spent a little time staring at it. This is interesting:



Red is cumulative % of total population in each zone (1991); black is cumulative % for Pakistani (2011).

So it's a basic indicator that Pakistanis are more concentrated than the population as a whole. But question: is that true for all groups? Is there something that sets than in stone for any subset of the total? I don't think so, no - logically, it's easy enough to imagine that e.g. zone x has 10% of total pop and 10% of pakistani pop.

Annoyed I don't have scotland in here. Might get all of the countries back - can subset afterwards.

I've just output a bunch of them. It's helping me think a little more clearly about it, but I still feel like I'm a way off. Probably enough for now.

..

So just thinking about one CoB again. Pakistan 2011 ~ 1991. Second log: has had 1991 zeroes taken out and ylim narrowed.

|  |  |
| --- | --- |
|  |  |

..

OK, so just looking at the numbers:

* For Pakistan, the shares have **dropped** in 2011 for the top valued zones. I'm pretty sure the numbers didn't drop, and I know you can see a growth in clustering in Glasgow. So I would very much like to:
* 1. Understand all the various patterns just for 1 CoB
* 2. Apply all that as a diagnostic to all of them.

So that requires going back to the original numbers to get some other numbers like zone % and raw numbers. Related question is: how much did population per zone change generally too? Spatial patterning of that?

And somewhere, dammit, I already had all the output for this... Oh yeah, found it.

So staring at 8nn's 11/01 correlations, an obvious point:

* Shares have the advantage of making different CoBs more comparable
* But their cumulative spread (as shown above for one example) is partly going to be a function of their actual numbers.

Comparison of cumul shares - as above - but between 2011/1991. Most CoBs became **more concentrated** (by this measure). Apparently Pakistan became **less-so**, even though I know there's some cluster growth. So I'm not really trusting these shares very much. Those are here:

**R\_outputs/shareComparisons91to11**

..

So the rationale for shares: if we think we're looking at 1000 people from Pakistan turning up in 2011, then shares tell us where they've gone. Problem: that's not what it's telling us. There's an annoying thing with the people already there and how they've spread around / not moved.

I suspect the raw number / raw percent partly determines the share shape as well. Small numbers overall mean more of a spread ... for reasons I'm slightly struggling to pin down.

I think... I should do a pairs for shares vs %. I wonder where that code was. Hmm, I can probably find it...

**censusDataLookz.R**

So let's compare shares and %s. Note: shares should/will be linearly wotsit to the raw numbers, right? % won't.

..

### Change?

Paws. So what about looking at change? As per Meng Le's suggestion? Something I can look at quickly / think through for obviousness?

If our questions are:

* Do migrants go where they already are?
* Is that dependent on who's in those zones already, or who's around?

Yes, that makes sense - so if we have shares (which I now see is only better than raw numbers cos it makes different CoBs comparable) to start with, and then see how those depend on **change in CoB numbers or share** between Censuses...

But and also but: it would still be good to know summat about which direction raw numbers have gone in. Err. Do I in fact have those?

Shares vs something like % change in numbers would make it comparable as well.

Now I'm doubting. I should probably have a look at one and see what I think.

..

Err. A lot of these zones have got **tiny numbers in**. % changes from 1 to 2 will be 100% - but not very meaningful.

..

Oh look! Geoff's paper has some references on this. Though I'm struggling with this. But we have:

* "Data on stocks of migrants hide the underlying patterns of gross inflows and outflows and differences in behaviour between cohorts. A constant stock could be consistent with high out-movements of existing residents, compensated by strong inflows of new migrant arrivals. Some insight into these issues is provided by Paccoud (2014) and Datu (2014). The former uses census information to examine differences in movements between 2001 and 2011 of migrants born abroad and of second generation migrants born in the UK. For most countries examined, relocation patterns of the two groups appear to be similar (Indians are the main exception), although the conclusion again depends on spatial level. The correlations are typically higher at the borough than output area scales. Datu considers information newly available in the 2011 census on the length of time migrants have been in the UK, which allows differences in the behaviour of the cohorts to be examined. From this, it is increasingly becoming clear that migrants are no longer confined to the Greater London region and are now spreading further into selected towns of the South East." [245/6]

..

So what are those two references?

### Year of arrival table

Ah ha. **QS801SC: year of arrival** - gives **all** years of arrival going right back. Which gives us a way to think about total numbers all the way through. There's (currently) no CoB / year of arrival breakdown, though we might be able to buy that table.

There's also this commissioned table for EEA countries with some breakdown that might help:

<http://www.scotlandscensus.gov.uk/documents/cde/CT_0071b_2011.xlsx>

Via:

<http://www.scotlandscensus.gov.uk/ods-web/data-warehouse.html> migration tab

..

OK, so now I've got that info I think I need to just discuss it next Monday, not do anything too detailed just yet. That table's very useful though - e.g. we've got a count of everyone who arrived in the UK since 1991. Not who's left or moved to other areas either within Scotland or England - but gives a rough idea of how numbers may have changed.

Good Lord, this hurts my head.

Well, note we have those arrival numbers for EEA countries... And actually, by comparing 2001/2011 CoB figures, that table might really help say something useful about how you can model this and what difference it makes when you **know** who's new and who isn't.

A lot of if/thens would be involved: shares of new count - assume no in/out flows, assume they're split according to existing shares. (Which begs the question, right?)

..

OK, so what I could do:

* Subtract those we know are **new arrivals** from the total of that CoB
* Work out shares of new arrivals - regress based on that, see how different it is.
  + Starting with correlation of new arrivals to overall count (as proportion?)

..

The Paccoud paper: subtracts e.g. CoB = Bangladesh from "Bangladeshi" ethnicity. The remainder (in theory) must have been born in the UK. Then what's it doing?

"Most of the increase in those born in a particular country at the OA level in this period can be assumed to have recently arrived in the UK." [29]

Well, can check that with the Scots data possibly.

..

### Ooo! Loo idea!

* So: if we have new arrivals for 2011: substract that from the 2011 CoB. That gives you everyone from that CoB who was already there.
* If you subtract **that** figure (the been-there-prior-to-2001 figure) from 2001, **you get the net internal inflow of people from that CoB**. Presuming the Census is reasonably accurate - if the numbers are large enough, it should show up.
* To test that, it might be wise to use reasonably large areas to illustrate the point.

I think that works. So for example:

* 500 in zone x from Poland
* 100 new arrivals 2001-2011 leaves 400 CoB in 2011 that have been around for longer.
* 300 Polish in ... ah dammit, **doesn't quite work**: we don't necessarily know how many new arrivals came in at that point in 2001 do we? Err, or do we? Err.

### Doing something simpler: 2001 / 2011 difference

So this is for the EEA table linked to above. I want to just test what difference it makes to the patterns for those countries that we can tell the difference from. Especially interesting for e.g. Polish - much larger numbers in 2011. Where did they go? Clustering? So.

Need to re-do the eViews thing. Which perhaps I could do with making... refactor blah blah. Maybe not right now. But what I need:

* 2011, I can mostly keep the same. I just need to add a split between <10 and > 10 years.
* I can also mark specific countries - I should probably restrict this exercise to those I actually have figures for, though it would be possible to get an "every other non-EU country" figure but not know where they are.
* Note: I'll have to make an assumption about ... oh, this buggers this up doesn't it? DAMMIT!
  + If I assume e.g. the Polish new-comers vs been-there-a-while are split the same in each zone: then I'm just **imposing** the clustering that's already there. It doesn't really tells us anything.
  + Except... I do **also** have year of arrival for all zones. So. So. Err. Can I do something about linking those two to make some kind of assumption? (Now we're into microsimulation right?)

That would require doing some sums with the year of arrival for 2011. Let's have a look-see. Note: **2001** **does not appear to have year of arrival table**. Checked pretty thoroughly.

## Urban / rural

So easy enough to attach this data (with some questions about overlap). I would then like to use it as an example to get my skin under interaction dummies. (Which is peesy to implement in R.)

First, though, getting data etc. (Rather more detail for Scotland.) Where was it again?

<http://www.gov.scot/Topics/Statistics/About/Methodology/UrbanRuralClassification>

OK, so which of those to use? Just the basic 2-cat? Well, let's try that first and see.

..

Huh: for some reason, the intersect in QGIS is being insanely slow. Something about the size of the individual urban-rural shapefiles I would presume. (Note this is urban-rural based on, what, 2014?)

Let's try R:

<http://gis.stackexchange.com/questions/140504/extracting-intersection-areas-in-r>

And Raster and see what we get.

Hmm: impossible to know what speed R is working at. Might be worth doing this with individual zones so see if we can do any better.

Just trying to run it on **one** of the two urban/rural zones (so then easy nuff to deduce for the other.) Being slow! Let's try the individual zone approach...

## More restating assumptions

OK, so going to Reading next week. Only a few days until the end of January, not much sign of a full draft. So a little bit of a list of things to do. (Having spent the morning recovering files that deep freeze deleted. Think I've probably got most of what I needed...)

* Get some chunks of writing done. 45 min slots of timed bursts to get moving on it.
* Complete 5 census data assembly. Issue of house prices / dwelling count.
* Do something, not sure what, with the urban/rural data.
* Dig into "what do the numbers mean" some more (as doing with shares).
* Related: dig more into regression generally. Regression101 in learning\_stats has some good synth stuff on dummies - I can see why it only shifts intercepts now.

### Bit of writing #1: PCS as common zone

It's pretty stilted at the moment and I want to account for having shifted to purely PCS based. Which may mean re-writing everything that's already there. It's hardly a huge amount so maybe not a big deal. But let's just have a go at informally writing it here to trick myself into writing it properly.

So why the 19 PCS? Well. Two-fold: to get a consistent geography for both the 5 census and 3 census data to make it as comparable as possible. But the choice of 91 PCS itself? Dictated by 1991 Census LBS table for country of birth - it has 90 categories, making for a much richer list of countries that can be harmonised from 1991 to 2011. But in order to capitalise on this richness, the PCS zones must be used.

They also need some tweaking: the zero-zone issue. So yes, this is more the way to write this.

..

The data is harmonised across time both geographically and across country of birth categories.

### #2: can I get on with background?

Maybe?

### Country of Birth context

Just some basic numbers. Scots aggregate count will help to take a guess at how many might be pre-existing. But **it won't help with internal migration numbers**. So I might just check for WICID flows at the national level - can it tell us anything about non-UK born into and out of Scotland?

Though of course those flows will be different for different decades. Ahaha. But we could try and find out about 2011 just for illustration porpoises.

First WICID attempt:

* We only have Scotland as a destination, not an origin. So it's only from English data out to where people were the previous year. (Not from Scotland to England, for instance.)

..

Oh look, there's something specific to Scotland. Is it any good?

<http://www.scotlandscensus.gov.uk/origin-destination-statistics>

Which links back to this, which is surely going to take me back to the WICID page shortly...

<https://www.ons.gov.uk/census/2011census/2011censusdata/originanddestinationdata>

Yup, WICID. Urgh. So what the hell?? Well there's this PDF guide, so let's see:

<http://www.ons.gov.uk/ons/guide-method/census/2011/census-data/2011-census-prospectus/release-plans-for-2011-census-statistics/subsequent-releases-of-specialist-products/flow-data/origin-destination-data--user-guide.pdf>

That has a good guide to the codes used in the different tables.

..

OK: <https://www.nomisweb.co.uk/published/census/odexplorer.asp>

OD by CoB / sex - **secure setting only**. Shame they don’t have it just by CoB / country.

..

Ah, might be getting somewhere. "MF02UK - Origin and destination of international migrants (expanded country of origin)".

p.s. here:

**C:\Users\admin\Dropbox\SheffieldMethodsInstitute\Census\_dx\flowData\2011\OD\_migrants\_MF02UK\_nomis.csv**

OK, so what that's given us:

* Numbers of people in each LA who were abroad a year previously (2010).

Codes can be summed to country easily enough. But could also use it to check variability compared to LA CoB sums. CoB categories are again different I think. But we have a simple **outside UK a year ago** figure here.

Which means. Well... not a huge amount really. We don't know who's arrived in the previous **ten years** since the last census.

Well OK: might be time to pull out of this rabbit-hole and do something else. Like check I've got everything together for 🡪

### Back to the 5-census data

I was getting it into an eViews-friendly format next, wasn't I? Did I also give it sensible names? Err. Yeah, fairly. OK, so. Given CoB\_combo some sensible names. Now to make the list of countries. (Tempted to add the full list back in to the 3 census one too - can then reduce to the smaller number. More work though, of course.)

..

OK, so having left it an eesny bit late, I could do with just knocking this together to show regressions from it. So.

Note: if I have time, it would be good to break down the "arrival year" numbers - so I can get a guess at what we might be able to say about:

* What would happen to housing and employment markets with a sudden hard brexit? Losing x, y, z numbers? Housing markets in particular, and some estimate of local econ impact?

OK OK. Sooo. Got some basic numbers. Want to do anything exciting with them at this point? And what else?

Well that's probably enough for now. Not sure I feel terribly... anyway.

## Geoff visit bits [30.2.17]

Paper: Gary Craig / superdiversity - "migration and integration"

..

Model: **effect of education**, interaction with CoB: are some more excluded from labour market, does this affect their clustering?

..

Glasgow visit:

* We need to show data output generation ...
* Some simulation bits and bobs

..

* Paper: working out what the theme is.

Instruments:

* Use 2001 as instrument against 1991? For all vars?

Jobs:

* Run urban/rural for urban only for cities in Scotland for poor/rich. (Urban only - cf. highly sig for urban.)

Urban/rural:

Identify areas where lag-dependent variables are **not significant**.

Are coefficients homogenous across the two groups?

Run across diff cities

Q: does number of migrants in cities = greater spatial persistence?

Are there things that are very stable? How does that compared to things that change?

Industrial change -> if we're extending to GB, gets more interesting.

If dominated by urban - are some towns different? Are there some specific shocks that account for those?

Or do they all behave in similar ways?

That gives us a reason to look back to 1971 ... looking for other shocks. Narrative. Are there differences?

Thatcher effect? North/South divide.

Then link to accession.

..

Employment: is there some kind of exclusion of certain groups locally? Employment as function of education / social class.

"Employment integration index".

..

Spatial persistence interaction with the neighbourhood itself?

Employment: what happens w/ internal migration?

Another literature: factors affecting national migration.

..

Different distances determined by different factors: short-distance = housing, longer = employment.

Do we have asylum seeker info? APS? LFS?

..

What to say about Accession countries?

Link to Casey review - integration question.

..

* + 1. Run separately on 4/5 scottish cities - urban / rich vs poor (are diffs across 4 cities?)
    2. Run on whole Eng/Wales/Scotland, same thing (3 weeks)
    3. Run separately on 10 largest cities in Eng/Wales (ttwa? Or just largest?)
       - Commonly used acceptable def of city.
       - Sensitivity analysis: re-run based on removing zones incrementally from edge
       - [employment differences?]
    4. Accession country breakdown
       - Do migrants from accession countries have different spatial persistence?

If there's no diffs between areas across areas, can the 5 census data show there's a longer-term effect to be picked up?

Unit root?

..

Some theories (Gwilym):

* Larger cities = ^^ spatial persistence due to sorting / size effect / path depedence following on
* Add in distance to countries: closer ones, arrivals = less need to connect back to country of origin. So theory: distance to country positive correlation to spatial persistence.
  + Which would be a way to cross-examine where rich/poor is due to networks. (Or showing that there are a couple of theories to explain it.)
* Cultural distance: accession countries geographically but not necessarily culturally proximate
  + Matrix of CoBs in own zone and contiguity. Correlation?

## Notes of the notes above

For sending round. So plus whatever's in the notebook.

...

Dan's tasks:

* Run separately on 4/5 Scottish cities -
  + Having already narrowed down to "urban" via rural/urban analysis, then do rich vs poor. What are differences across those 4/5 cities?
* Run on whole Eng/Wales/Scotland, same thing (3 weeks for getting England/Wales data together)
  + Run separately on 10 largest cities in Eng/Wales, look for commonly used acceptable definition of city.
  + Sensitivity analysis: re-run based on removing zones incrementally from edge
  + Employment differences?
* Accession country breakdown
  + Do migrants from accession countries have different spatial persistence?
  + Link to Casey review / integration question.

If looking at urban only: are some towns different? Are there some specific shocks that account for those differences? What things are stable, what things change? Thatcher effect, North-South divide.

Or do they all behave in similar ways? Either way: offers a possible narrative reason to then use 5 census analysis to look further back to 1971. E.g. if there are no differences between areas for some value, can the 5 census data show longer-term effects? (Possibly! Maybe different paper.)

### Other model ideas

* Effect of education: are people from certain CoBs more excluded from the labour market than education levels would predict? Does this affect clustering?
  + Employment as a function of education / social class: "employment integration index".
* Size effects:
  + Does a larger number of migrants in cities mean greater spatial persistence?
  + Do Larger cities = higher spatial persistence due to sorting / size effect / path depedence?
  + Which would be a way to cross-examine where rich/poor is due to networks. (Or showing that there are a couple of theories to explain it.)
* Examine different distance effects. Different distances determined by different factors: short-distance = housing, longer = employment.
  + Examine distance between countries: closer ones, arrivals = less need to connect back to country of origin. So theory: distance to country positive correlation to spatial persistence.
  + Cultural distance: accession countries geographically but not necessarily culturally proximate
  + Matrix of CoBs in own zone and contiguity. Correlation?

# February 2017

## On with migration

So a nice break after all that. And the first thing to do:

* Having reduced the sample to Scots urban, re-run for the separate TTWAs.
  + And for rich/poor within those cities.
  + A good chance to look a little closer at the actual patterns. Though note an issue: **W will include edge cases** so will not be quite as sensible?

Just checked the TTWA selection is doing what I think it is. Yup.

..

OK, doing some urban regressions. Just checking that people are going where pop share is. Now... pop share was total population share **including all CoBs**, right?

Yeah. Just done some plotting: predictable TTWA-level correlation overall. Not quite sure what to think about that at the moment.

Well, I should probably just get those things done that need doing!

..

Err. Done that, sort of. So I have that output. Have to decide what to do next now I have it.

### ­Working out shares

This is ever-so more complicated by thinking about comparing cities or other groups. If they sum to 100 over the whole group, trying to correlate a subset of shares is going to be confusing.

..

Random thing on regression line vs correlation, which is just supporting my confusion... though...

<http://stats.stackexchange.com/questions/32464/how-does-the-correlation-coefficient-differ-from-regression-slope>

And: was just wondering why flipping variables didn't reverse the line polarity (i.e. 0.75 being a mirror on the other side of 1.) Remembered: **least squares**. It's squaring y values / along the vertical axis.

So the coefficients are... more or less arbitrary then? (What about robust regression?)

..

Let's just get output for each CoB to see what the variance is between them. Noting the problem of shares again, but this is perhaps less of an issue here.

Related to the above, which is possibly what I need to be thinking about too:

<http://stats.stackexchange.com/questions/78640/assessing-fit-with-identity-line-in-q-q-plot>

..

So I need to decide what to do about all this at some point soon.

### Braindump from staring at CoB 91/01 regression data

In:

**Dropbox\SheffieldMethodsInstitute\Census\_DX\R\_outputs\urbanPlots\urbanCoB**

..

So this is for urban zones only. And staring at them I want to ask Gwilym some things. So:

* More evenly spread countries tend to be the whiter / richer ones. They tend to be much more linear just because they correlate to the underlying zone population. So e.g. look at Canada / Australia. (Smaller max share = population is more spread out.)
* Others with higher concentrations: line is weighed by many datapoints at origin.

### Randomising

Just trying to come up with the random distribution of people. Which is proving annoying. This might be promising:

<http://stats.stackexchange.com/questions/26858/how-to-generate-numbers-based-on-an-arbitrary-discrete-distribution>

Oh look, vector of equally spaced values - exactly what I've been doing. Or... ah, a lovely easy sample way of doing it. Knew it had to be there somewhere. But:

* What things to compare to?

I think this is right, but there's some other things to think about:

* What about if we presume x CoB would move ... no, that's the thing we want to show does happen. But how to show that by comparing to the existing population? That is - if it's significant (because values are beyond what we'd expect from average assignments), how to then show that it relates to where their own CoB are?
* Same applies to showing there's a **spatial** relation to surrounding zones.

p.s. <https://en.wikipedia.org/wiki/Poisson_point_process>

So point process in Euclidean space: the assignment to zones by area is going to be analogous to what I've just done using sampling by probability. I think. So yeah - Poisson process, might help.

### Aside

Back from meeting. Managed to get over fearoflookingdumb and ask some really dumb questions, turns out Gwilym isn't completely sure either. So need to draft something to ask Geoff on this. As well as get on with prepping the GB data. And lord I could do with tidying the code, or in some way doing it all in one go - given it will all need to be linked anyway.

Err. So what order do I want to do things in now? There's rather a lot to get done. I would like to nail down a bit more what I think the stochastic processes are (now I have a simple start on that.) Think I might need a walk and some snack buying first.

..

But now I'd like to try and think about the probabilities again. There's a slight worry it might get a bit bayes. But I don't know.

I still can't think of a way to approach it. I suppose I could investigate by just seeing how each CoB differs from what we'd expect given averages. That might be a start. Or a pointless faff. Hmm - let's see.

..

Well, my conclusion from all that is, the distributions are the same, more or less, when turned to shares, regardless of number of people. Which I guess they would be.

So that's probably enough nonsense for one night.

### More on what the share regressions mean

So currently staring at America again. Hmm - it was the one example where the regression line was more than 1 - but it was urban, and a chunk of US folk were in rural areas. Re-running for the full 100% share, none of the lines are above one. Is that really not possible?

Oh hang on, I think I've just worked out that it's not:

* Shares are the equivalent of normalised variables - **so this is the equivalent to R-squared.** (Which I'd already read about.)
* So to understand it, I need to more clearly understand that.

OK, so back to those links about covariance / correlation etc. Some more links following up includes this:

<http://stats.stackexchange.com/questions/22718/what-is-the-difference-between-linear-regression-on-y-with-x-and-x-with-y/22721#22721>

This answer is great but another question I now need to understand:

* Slope is the same as correlation (Pearsons) when both variables have been **standardised** (subtract mean, divide diffs by SD).
* But we're working with shares. Not quite the same thing, but it might end up being the same thing mathematically (or nearly - I sense it's a little bit different. I can probably show that through randomisation.)

I think I need to get under the skin of basic covariance shizzle before this will make sense. Possibly. Which might mean doing what I've just done. This is before we get on to adding in W and thinking about the collinearity involved.

Holding off on proving ('prove' through brute force) that it actually can't be more than one. I don't think I need to: I've reversed regressions between years, it hasn't changed the line from being below one.

..

Having just reversed it, it is - **just** - possible for the coeff to be more than one. Which presumably it couldn't be if using covariance for R^2. So yes: shares are similar but different.

So Pakistan is the one that's slightly more than one.

..

A pause for a question:

* **When, if ever, would it make sense to recalculate shares for subset areas?**
  + **Is it OK to redo shares just for urban?**
  + **What about comparing cities?**

Like when comparing cities? The problem if you **don't**: you're not comparing like with like. The results (as America if excluding urban shows) end up being warped by places where numbers were very different. So it only really makes sense to compare 100% shares with the same.

That said, there may be a way to dig out the meaning of comparative share changes if e.g. Glasgow overall gains from somewhere else. (Again, this is sounding hierarchical and complex!)

But does it make sense to re-do the shares? So something like:

* Of those that **do** move to Glasgow, how does that determine where new arrivals go? (Putting aside that we don't know if they are new.)

OK, compare these two. This is something very obvious I hadn't really mulled:

|  |  |
| --- | --- |
|  |  |

Intercepts: 1.75 on left, closer to 1 on the right. Obvious thing:

* Right-hand side has **100 values top right**, left has **one**.
* So when proportions are done, each datapoint on the RHS ends up with much smaller shares (despite being each of the same size - just more zones.)

So need to stick to the same number of zones to make sense of this. At any rate: that's getting values higher than one for same-zone-number.

This is also important. We knew this but worth seeing. So:

* The first 10 points stay the same: 1:10. They would usually therefore count as unchanged.
* But:

|  |  |
| --- | --- |
| B = 1.27. In top 10, numbers increased from 20 to 40 (so reduced shares relatively in the lower 10.) | B = 0.76. Numbers in top dropped from 40 to 20. So **relative shares** for others in 2011 is higher. |

|  |  |
| --- | --- |
| B = 1.7. Top stays same (w error), bottom **drops** from 11:20 to 1:10. Top shares inc relatively | B = 0.58. Top same, bottom **up** from 1:10 to 11:20. Top shares dec relatively. |

OK, so that's useful. It means the coefficient can have the same values while meaning something different. Which is confusing. And in no way am I seeing how it can be a comparable measure of something like spatial persistence.

For a start: change-wise, it means:

* **Any that stay the same while other increase will be pushed into the <1 side of the line**.
* **In terms of the slope coefficient**, the difference in impact will appear opposite depending on where in the rankings this change happens.
  + If large-zone numbers increase, small-zone ones will be pushed < 1-slope, B will be > 1. (I think this is what happened with Pakistan.)
  + If small-zone numbers increase, other way round: beta will be < 1.

Or to put another way:

* From beta, you can't tell an increase in small-zone populations from a decrease in large-zone populations.

So what might be useful:

* Examine the **actual** zone population changes and check for differences in patterns.

OK, got all that graphed. Staring at it not very useful: I need some way to pick out CoBs with differing patterns of change.

## England/Wales again

Haven't entirely settled my issues with shares yet but want to feel like I'm making progress with getting the rest of the data together.

Just looking again at area comparing PCS and wards: wards are overall smaller **but** I think that might change if we look just at urban areas.

Quick check in QGIS: Scots PCSs are more variable in area, I think. I suppose I'd need to check variability of population.

For now, though, on with aggregating... And I'm not sure where I got with this. And I am sure it's confusing as hell. Reminder plz.

OK, the compulsory restating of assumptions:

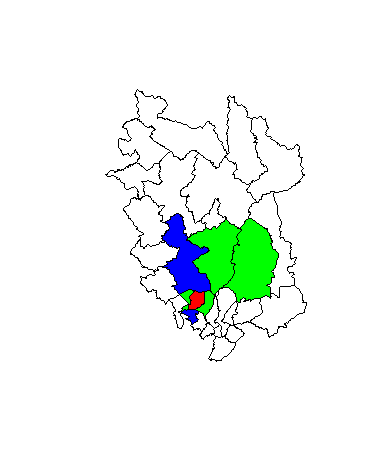
* If x ward has a population of zero\*:
  + EDs within that ward won't\* - they'll have the count.
  + Those EDs: their count will, mostly, match what's been added to a neighbouring ward. So look for a neighbouring ward with that number of extra people.

\*Mostly that works. Except for:

* Some zones will have more than one zero-set ward added to them.
* Some of the numbers here are less precise than for Scotland.
* Some of the wards are **actually** zero: but for those, the ED count will **also** be zero. (E.g. central London, no-one lives there. Apparently.)

### Pairs and more

It's slowly starting to make sense. I really need to blog this one at some point so I don't completely lose the plot later. But the pair matching test is getting there. Here's the current example: green are zero-set, blue are candidate zones for one pair, red is a successful match.



Unfortunately... hardly any pairs match. Of those that are pairs. There are only 35 left, it's tempting to try something else, but... let's looksee.

After all that:

* Only three pair matches. Oh, well, that's six more zones out of play.
  + Wait, no, it must be three. It should find both pairs. HOW CAN IT BE THREE??
* I suspect there will now be a lot ...

Ooooh wait. I've only checked **one of any of the list of pairs.** Though I've just printed out for 28,29,30 and it's finding three "pairs" equaling one zone amount, which I'm finding hard to believe. So need to work that out before doing anything else.

**Oh, actually, no, I think it's doing its job: for any set of possible pairs, it's finding the same one several times. Makes sense.**

So now I just need to make sure it loops through all possible pairs.

..

OK, that got **three effing pairs**. I was worried it was dumb coding, which it is slightly, but not for the reason I thought: it finds the same pairs repeatedly, but I think it needs to because it needs to get zones that are 2nd order neighbours.

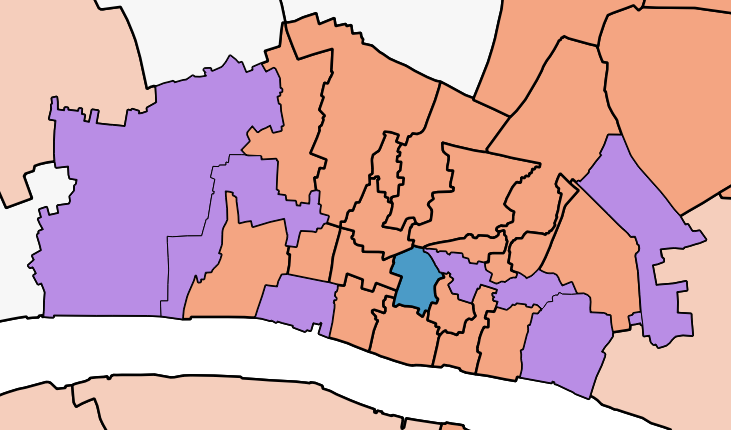
So fine, if a ridiculous way of getting three pairs.

..

Now onto the rest. Took care of the Isles of Scilly, they're all combined. Now time to look through them one at a time. Most look like single matches. Thusly...

Right ee ho: narrowing it down.

London: these six going into that centre one. Not exactly the right numbers but close enough given the various EDs pushed around here. Obvious this is where they've gone.



All but two are now straightforward. I'll look at those tomorrow. Note also: that London zone may need combining with some others just to make it a sensible / contiguous shape.

A lot of the zones around there have population zero: they can all be combined, most likely.

Actually, I've just realised:

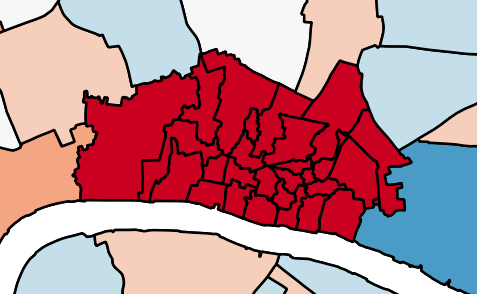
* It's probably the case that in those double-zero (ward/ED) zones, **both** the LBS ward and ED values have been assigned elsewhere. Certainly in that part of London, and that tallies with what's going on in surrounding zones. So yeah: that whole area is probably good for collating.

Though something that's just occurred to me:

* In places where there **is** a slight difference, do we need to look to nab the re-assigned EDs back into the correct zones? Otherwise pop bases are going to be wrong? Hmm. Will look tomorrow.
* Update: no. I realised this walking away. The underlying LBS numbers will be correct, it's fine. Phew.

So let's check on London first - I suspect many double zero-zones should be stuck into a whole with Walbrook.

The red are all SAS is close to zero or zero. Likely candidates for combining:



Let's just think through any problems with combining. I'm pretty sure if any of the LBS values had been re-assigned elsewhere, we'd see higher numbers there.

Having stared at it for a while with some additions: there are some quite large diff numbers in the rest of London, but I think nothing that calls into question the basic assumptions. Zero LBS zones are still what they are. There don't appear to be any zero SAS zones explaining some of the differences, which is a slightly worrying thing.

OK, I'm gonna go with my theory that those are all going into Walbrook.

Update:

* All Central London ward codes starting 01AA are to be combined, apart from Cripplegate (which is the north / central / Africa-shaped zone above) which has its own correct pop.
  + Oh and Aldersgate, west of Cripplegate.

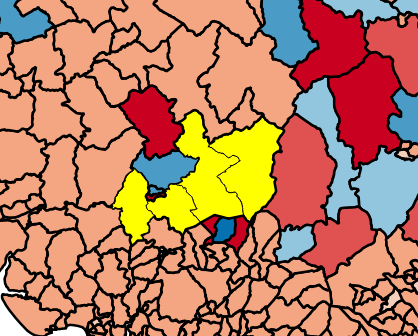
### Last two

And they're confusing. I think I need to be able to see which zones have **not** been taken (if any). So remind me what I'm after when doing that.

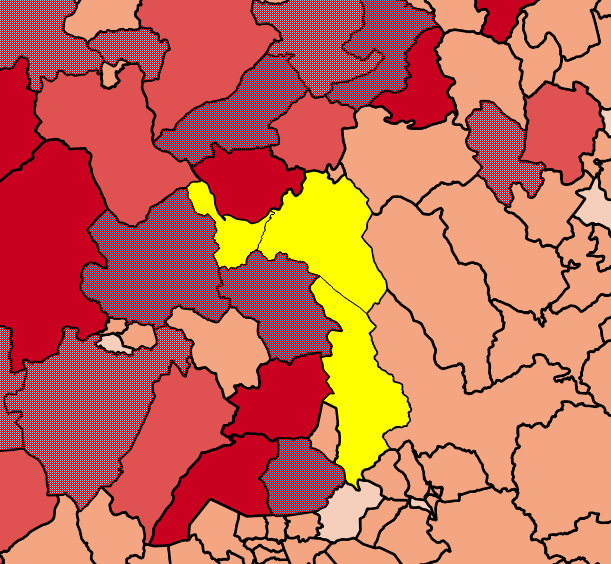
Err. So I need...

..

OK, found Myddai's home. That's odd - it's not contiguous and I can see no reason why. There are two more zones in the way. Only solution (since we need to make contigous zones): join all four together. It'll be big. Right-most is the zero-zone, left-most where it went.



Same for Bronllys (top of these): bottom is where it went, need to combine these three.



### And what now?

Got them all, or as good as I can get. Now I have to remind myself how the next bits of the process works. A couple of separate things to do:

* Create the new combined shapefile. Then for putting through intersect. (So will need intersected geogs for that.)
* Re-aggregate existing 1991 data to those new zones.

I did wonder about combing with Scots data and running the whole thing together but I think that’s the wrong way to do this quickly. I can just produce all outputs for Eng/Wales **then** combine - perhaps after I've got the "raw" shapefiles done, rather than later.

(Noting I'll have to do this for 5-census too, but let's get on with 3 first? Urgh, it's a tangle and no mistake!)

..

It appears to be here: had to do find in file search for where the scots 91 PCS was saved:

**Census\_DX/tests/1991\_pseudoPostcodeSectorChecking.R**

Let's just check it is definitely here... Yup, that's definitely it. And all we're working towards: **unionSpatialPolygons** on what are now same-ID zones. Let's just transfer the code across and do it in the same place.

..

Looking good, after one fix. One last job tomorrow: attach the labels for zones that have been merged into the one that remains.

Done. So now onto ... stuff that needs a new heading!

### Getting England/Wales data together

Would like to just run through again what I'm aiming for here. A little planning would be good given the mess of different things. Recall: I'm going to aim for a separate Eng-Wales data bin then link with Scots afterwards.

Then eViewsReady and all that: I can do that on the combined GB data. Though I want to just do a little comparing of zone sizes / populations etc.

Right, so, here's what we're after:

* 3-census: Country of Birth / Dwellings / Employment.
* 91: each aggregated to the new zones.
* 01/11: each put through the pre-intersected geographies.

So I need to check what data for England / Wales I already have. And possibly get those geographies too. Then intersect.

* In EngWales 1991 folder so far: totalPop (used for working out the zone combining). Do I have other data elswhere?
* Hmm. No. Which is going to mean some further processing, especially for country of birth. (Male / female adding up, that kind of shizzle.)

Let's try and break this down a little more neatly: get it all into reasonable form **before** any further aggregating. So where? Err. Start with download, dude.

### England/wales CoB 91-11 download / wrangle

What script was I using before? Is it a sensible place to carry on? No, don't think so. Urgh. Yeah, new script:

**englandWalesDataWrangling.R**

..

First, downloading.

OK, so this is interesting. I haven't actually had to download 2011 CoB England / Wales data before. And I think it's possibly fucked. InFuse is awful. It's difficult to know quite what we've got. This is after having to manually select each of ~330 categories. I'm hoping somewhere in there we'll have everything we need.

It appears I can get 78 categories from NOMIS:

<https://www.nomisweb.co.uk/census/2011/qs203ew>

But only by region. 78 might work depending on what those categories are. Note: England in InFuse appears to have eff all. (Wales is, possibly, fine. Will need to look.)

Note also this list of commissioned tables. Not what we're after though.

<http://webarchive.nationalarchives.gov.uk/20160105160709/http://www.ons.gov.uk/ons/about-ons/business-transparency/freedom-of-information/what-can-i-request/published-ad-hoc-data/census/ethnicity--identity--language-and-religion--eilr-/index.html>

..

So, check through NOMIS: we've lost a few more categories but nothing too major. We still have most. But it does mean a re-combining for all the other data too. So, err... woohoo!

It's on this printout here, so will mull what to do next. Also double-checking there are really no more detailed categories anywhere.

Just looking at InFuse's **UK-harmonised Country of Birth**: I think the categories look the same as for the NOMIS bulk download.

OK, yes: that's populated. Which is probably good as it means I can download for entire countries at the same time.

In fact: I can get England / Wales / Scotland at the same time. But that would then mean having to do some arsing about with Scots data again which I'd probably rather avoid. I should be able to throw those few countries into "other" for Scot. Well, let's check tomorrow!

..

OK, back again. So I think it's going to be easier just now to tweak the existing Scots CoB data to match this new Eng-Wales list. Things to note:

* **New Zealand** to "rest of world" cos "Australasia other" includes a bunch of smaller islands, not just NZ. (Though the difference is obviously going to be minimal.)
* Same for Malaysia and Singapore.
* Canada: grr. Same. Too many little Islands.
* Greece, Netherlands: other Europe.
* Note, original category that appears just sa 'Channel\_Is' - this is "plus Isle of Man". So that still works for combining.

So that's just a little recoding of Scots CoB. If I do this all in a GB folder I should be able to keep it neatish.

..

OooooK. **It would appear that the InFuse "Country of Birth (harmonised)" list is missing some countries that NOMIS doesn't.** Let's just check I can get everything in NOMIS. This is going to involve downloading each region separately.

For Wales at least: yes, we have all those categories. So InFuse has less. For some reason.

OK, need to: download all Eng-Wales regions separately, check each has positive values / no blank columns. Wales is good.

Let's just check that NOMIS is giving us the correct geography... yup, looks good.

And the numbers check out. Bonza. There are even 49 people apparently born in Antarctica!

OK, saved the combo file for 2011 CoB. Off to mum's, will work out next steps in a moment...

...

One last job before finishing for the day: combine 1991 and 2001 data into Eng+Wales.

OK, well, the basics done. I have to add sensible names at some point, though I can't remember what the optimal point to do that was. (And whether my existing categories from Scotland will match. They should, right?)

* **"Oh dammit: 2001 - non-EU countries in Western Europe":** I think I deleted it as a sum column when it's not. (In Scotland.) It likely needs summing into "other Europe."
  + It's possibly a reasonably big number. Grr. How much of Scottish data does this require fixing? GRRRR. Can I just run it for these numbers and add them? Will mull.
  + I could do with adding it on the end of the existing dataset, otherwise I have to change all the CoB stitching reassign numbers.

OK, so I have names updated for 91 and 01. (With the previously missing other-W-europe category now in.) Just need to reduce the 11 columns to the actual countries. Easier doing that manually. Enough for now!

### Where we at

Two weeks until Glasgow meeting. I've got a list of other viz to do that gwilym gave me (though not the data yet). What I think I'll need to do (to say to G today):

* CoB almost done. Get some viz / info out from that once we have it. (Contiguity matrix doable then. Though: will need to link Scot / Eng / Wales for full effect.)
  + Don't worry about other variables before...
* Gwilym's list of viz: which to do first?
  + There are three (point five) so discuss at meeting.

Right now: get some GB CoB info done. Done. Next:

* Stitching these 3 CoB for Eng/Wales into common categories.
* Doing the same for Scotland with those two other countries changed...
  + Oh so I did need to do this anyway. So altering the awry europe category no biggie I guess.

Err. What first? Well I need to make the intersection files too. But stitching... happens... before that? Right? Yeah. OK. Might as well add to the same 3-census file.

In theory I should be able to work this out with less faff. We'll see. Err. Maybe not! Maybe need to print ... but can print those that don't already have a match.

### Some notes on the EngWales 3 census COB stitching

* Jamaica seems to appear as its own country in each. I didn't think that was the case in the Scots version. So look out for that when combining the two.
  + Hang on, can't I just check this?

### Emailing Geoff about the coeff (geoff coeff!)

Been putting this off, need to just write the fucking thing and send. So what am I asking? Remind myself of some of the details.

Oh, one of the things:

* Should we be recalculating shares overall for subsets of urban / poor / rich? Without that, you can end up with some funny numbers. Need to be clear in my head on that.
* Looking at notes above:
  + Regression line can depend on number of zones and their position: 1 large zone top right will have a bigger share than 10.
  + This might not be a point worth making right now, but worth remembering.
* On the randomness tests (bulk-regressing random shares against each other): I've also just compared 1991 actual to 2011 randomised and the values for SD are lower. Probably as you'd expect.
  + But the total randomness also depends on the number of "zones". Again, this seems to make sense when I think about it (smaller numbers will mean each number can have a higher variance when converted to shares - it doesn't matter what their original size is.)
  + So I suspect this is quite different to normal significance testing. But I don't think I have time to get into this. Can I find some way to mention it?

..

So getting on to the more important bit: what the number really means. Looking again at what I've got. This is America, all then urban. So a couple of things: urban = not re-calculating shares. So the numbers are whack. Possibly.

|  |  |
| --- | --- |
| C:\Users\admin\Dropbox\SheffieldMethodsInstitute\Census_dx\R_outputs\urbanPlots\CoBall\United_Sta_0.0949.jpg | C:\Users\SMI2\Dropbox\SheffieldMethodsInstitute\Census_DX\R_outputs\urbanPlots\urbanCoB\United_Sta_1.4156.jpg |

I still need to effing send this. I can't drag it on for two days, that would be insane. So. What else to put in?

Sent! Onto the next thing.

### For Glasgow: coeff bar charts

As requested. And I believe to compare to Geoff's too. Though yeah... waiting to see what it actually means. In the meantime, as requested.

Geoff's paper only has "all countries" and "europe" for the normal estimation.

There's a rich vs poor IV estimation but. Yeah. Err.

Let's get our own set up before attempting owt else. Urban broken down by TTWA will be tricky in Scotland. Or if that's too boring right now, I should do something else! I think it might be.

..

## Finishing off GB data

This is all being done in one R script thus far:

**englandWalesDataWrangling.R**

..

Actually quite a lot to do: "finishing off" probably a rather optimistic view. Sooo. Let's look at where we got to. Err.

* We've got the aggregated ward shapefile (annoyingly different from PCSs).
* I already did a bunch of work on stitching CoB right? And there are some changes that need applying to Scotland. And there was another error in Scotland from the previous combine. Which was? See **"oh dammit: 2001 - non-EU countries in Western Europe"**.
* I was mulling whether to do EngWales stitching separately then combine with Scotland later. I think it's probably better to combine the data beforehand into various GB files **then** do the stitching.

..

Soooo I think I need to restate assumptions.

* Create single GB data files for CoB, employment, housing stats.
  + 1991/2001/2011: higher-res CoB based on Eng/Wales CoB limits.
  + 5-census I haven't yet looked at.
* Whack the CoB through stitching to get same 91/01/11 categories.
* **Then** bin into common geog.
* And bin the other vars.

..

OK, that seems like a plan. Hmm. Except. It needs another dimension:

* Scots vs Eng/Wales CoB: needs same categories **in each year**. So that's the first job.

### Job 1: GB, 1 datasheet per time period.

CoB to start with: let's check what the categories are and how they differ. They're probably small enough in number to combine manually. So need to just look at what we've currently got and compare.

OK, first odd note:

* No category for "north africa" in scots 11. So will need to stick that in "other".

..

Got a list. Let's see about making both into the right thing.

..

OK, now I'm just checking matches for any names not containing "other". Should have all ones to combine, in theory. Haven't run it yet, will do that next.

Being silly: if I want to get matching columns...

Christ. So I now have GB CoB for the three decades. Next... oh wait. First I need to check I didn't lose anything. So.

### Stitching GB CoB into consistent CoB categories across decades

Which will be an eensy bit different to before I think, just in some odd categories.

* Africa other: what's going in that? What does it **exclude**?
  + Kenya / Zimbabwe / Nigeria
  + North Africa has to go into "other"...? Because 2011 Scots and English don't contain enough cats to link it? Really? Yes, really. Not there. So we lose that into "Africa other", which I think I was doing before anyway.
    1. Would be interesting to know how the data with more African countries actually breaks down.
  + So remove those three from all, what's left is Africa-other? Oh, well... North Africa can go into that, can't it? **NO IT CAN'T! SEE PREVIOUS POINT! We lose it because of Scots/England 2011 mismatch**.

..

Oh good, I think I missed another one. If I can stick it on the end of 2011, shouldn't involve masses of recoding. "Other Eastern Asia."

Next:

* **EUROPE.** Who goes where, what can I make? Let's remove the exact-country matches first, see what's left.
  + Oh dammit: SPAIN. Right. Fine. (Had bit of extra text for EngWales...) Which has shifted all numbers after Spain one up. Oh perfect! So that's any above... 10. Well, that should be easy to change, right? Yeah. Phew. Right, on with the rest. Still all into Europe Other? We have a couple of other accession countries...
  + Six countries in each: France / Germany / Italy / Spain / Poland / Romania
  + Lithuania in 2011 may bugger "other accession". Let's see...
  + Hmm,yeah: I think there's no way of avoiding the rest being "other Europe". Well, maybe... No, I can get "other EU member countries" (non-accession) I think. Yup, that works.
* So that's it: single country matches / EU member states / other Europe.

Done. Next!

### Turning harmonising CoBs into harmonised Geograph... eeez.

Which involves a few things:

* 1991: direct re-coding to the new shapefile. Err.
  + Which involves having to make a new GB shapefile. But after that, hopefully the same code?
* 2001/2011: geography re-assign code.
  + For which we need intersects.
  + Which involves GB OA files for those decades too. So two more merges.

Let's do all the geog merges first.

* Join looks good enough for GB pcs/wards. There's no overlap, just the occasional gap - nothing that will have any noticeable impact on assignments.
* Oh, haven't actually downloaded England/Wales 01 geogs... err, really? Really. OK then.
  + So need to check about singleIDs etc. Woop! May need to dissolve by ID.

England: needed two separate shps, then to dissolve. Now we have. Now for wales. Which is probably a good time to stop actually. Then we'll see if there's any issue with scots / eng / wales boundary. Hey, let's just look... nah, good. Phew. OK, enough for now.

..

Ah of course: Welsh OAs a super-generalised or nothing at all. Just to be awkward. Ah, it looks fine: that's generalised. And a single file. So... all good! IDs all match as well, which is an unusual little surprise. Nice.

Now for 2011. Do I already have any of the files? England yes (not dissolve), Wales no.

Well that was **much** less painful. Getting there! What's next? Ah, the intersects. This will take a while... so far so good.

I should probably try and get the '91 data aggregated to the correct zone size. Where was that code?

### CoB: Aggregating 91

Well it was going smoothly. But there are a lot of zones still with zero values in. Better take a look at those. Though I'm struggling to remember what I'm looking for, hopefully it'll come back. Damn, it was nearly working then...

Ah - there are more zones in the CoB file than the PCS file. Let's just backtrack...

Phew: **all shipping**. (In the CoB variable file.) So all good. Onward!

### CoB: Intersect summing for 01 and 11

For which we need the intersect function and code...

Hmm. The size of 2011 may present an issue. Let's see. Just fitted. Phew - it shouldn't have **not** fitted. Something odd with memory there.

### Done... that bit

Right, so there it is. It's not, I think, the most amazing dataset ever: cities in Eng/Wales have nothing like the resolution of the PCSs in Glasgow. London is OK. Other places not so much.

So there's a whole bunch of other stuff to get done yet. I could start playing around with looking at the various maps, but that requires making pop sums and just want to pause and work out what's best...

Well, let's just add a pop column so's we can have an eensy quick look... OK, definitely better when we can make percentages. Looks OK, some useful things showing up I think. Might move on to something else now but let's make a rough list of stuff that needs doing next:

* Urban/rural defs
  + Which may not be the same for eng/wales. So.
* Other geographies attached: TTWA / country / region. TTWA plus urban/rural = city. Wards a bit crude for this.
* Employment / Dwelling count.
* Five census? Kind of lost track what to think about that. Bit of a bugger to recode for CoB but can have a look. (And I think we want it...)

Just trying to think... Yes, worth trying to get some stuff out.

This is quite a useful thing for thinking about it:

|  |  |
| --- | --- |
|  |  |

Pakistan as % of zone pop: wards vs OAs, in an area I know pretty well. So some indication of how crude it is at the ward level.

Right, time for something else...

### Onward

First thing: look for urban so I can tell G+G what the actual city resolution is (not great I suspect.) So where from and are the defs any good / how do they differ from Scotland's? And what were Scotland's?

Hmm, they're different, and the boundaries don't match mine anyway. I think a wiser idea would be:

* Density of people in each zone from the numbers directly.

Let's just look at that in QGIS, see what kind of numbers work. Hmm... areas with>750 people per km2 seems reasonable.

X people per KM makes it easy to do this in R. I can then compare Eng/Wales and Scotland for the ward / PCS difference more easily.

..

Oh Good. China seems to have gone missing. Was there a reason for that? I didn't think there should be.

OK, it's gb91: "China, people's republic of." Didn't match the other Chinas. So none of the column numbers change, it just **all needs re-running a-fucking-gain.**

Check united states as well. Yeah, doesn't seem to be in the list...

OK, that was a few hour's work just re-running to get China and the US in there. Now on to doing some regressions on what we've got to see what's what.

### Some eezee regressions

And getting the data ready for it. Which I think I'll just do in the same doc.

You know, might be worth mulling some functions for this. I want to do the following, some of which will change:

* Gather CoBs
* Reduce them to a subset of zones e.g. > 1000 people per KM2 **then** find proportions.
* Then do the regressions.

Well, getting there. Some surprisingly high proportion numbers but they're looking sensible on a map. Suggesting there may have been more concentration in 91 than now. But what are the numbers? Anyhoo...

Bored now! Just getting rich/poor together, slight mismatch in names from previous list.

Then I need to do the weights matrices. So shizzle for tomorrow I think. And then there's all the other data... wurgh.

## GB regressions / Scots paper

So going to take a little time gettng my head back into this. Some shizzle:

* Gwilym's suggested going back to the Scots paper first so that the method can be reasonably well written up. (And looking at shocks or summat. I'm still unclear on that.)
* I also want to get basic regressions re-run for GB county-of-birth before starting that.

Hmm. England rich/poor split is showing reverse pattern, though basically similar. Might mark London out and look. I suppose for these larger numbers of zones, you'd expect more stable patterns. It might break down at individual city level - but not really enough data points to break down. Could do groups - as Gwilym suggested.

Hmm again: think it might also be worth **looking only at zones with neighbours of neighbours**. There are a lot of isolated zones currently.

Oh also: it could be because I've done the proportions based on only urban zones. That could be... it, actually. I wonder if I did that with Scotland, the same would happen?

* Nope: it's actually higher coeff for rich now.

..

### Neighbours of neighbours

Checking on count. So got the first queen contig. What next?

OK, done: it's not perfect as some neighbours of neigbours counts mean single zones in "islands" get cut off. I think it could do with... would this be easy? Err:

* Any zone that has a neighbour who **did** succeed the "has neighbours of neighbours" test should be included. No, that doesn't quite work either, though it's better. Will come back to this!

Right... no, that made no effing difference. Tell you what though: I could do with functioning up this code, it's a pain in da arrrz.

..

So while I'm waiting for my brain to clear, can I just check, did I update Scots CoB data directly? I... err, no I don't think I did. I will need to. Errrrr.

### Functionalz

Should speed things up a lot. Actually... the simpler solution is to just not do any...

Oh no: the point is **keeping 100% proportions** for whatever subset of zones is used. So that's zone subsetting, not CoB. (Each CoB will be consistent.)

Let's just read through the code and figure out what to put where. (In englandWalesDataWrangling still.)

..

Scotland is giving different results. Think I'd better just directly compare the numbers... no, they're all fine (more or less). Some changes from different binning but fine. So it's not that. Must be addition of weights matrix? OK, so those need creating.

### Weights matrices

Need these! So so so. So done after some faff. And we have curious numbers. So before putting them all together... no, actually, I will put these together. Then I want to compare to e.g. Scotland again now that we have the weights matrices.

..

Still odd numbers. Let's just try: for Scotland, keeping proportions for all Scotland but removing zones **after** the proportions are calculated.

### Issues

These numbers are looking too different to previously. Need to go through carefully and see if there's any reason why and do some direct comparisons where I can.

* I've already checked the underlying numbers are as good a match as they should be given different cats. Has something else slipped up later? Actually I can probably check that directly.
* Changing to the original urban zones doesn't fix it. (If indeed it needs fixing.)

Ah - I realise one thing we haven't checked: **using Geoff's subset of countries.** What were those?

OK, that does make a difference, though perhaps not as big a difference. I should try it on GB though...

## Some restating of assumptions

So I need to systematically work out what's going on with any data differences. I tried a lot of things yesterday but it's all messy and I got lost with it. I need to compare directly to the same previous Scots data, for the same countries, get zones right, splits right etc. And compare other things - differences in urban / rural as well as finding a way to keep only groups of zones. (I think I worked out how to do that.)

### Functionalism again

So I've tidied it once but it needs tidying some more. I need to be able to load and compare both sets of data.

Actually, I should also just save what I've got for GB as maps and check the weights matrix has worked. Let's do that first... **Yeah, that looks fine.** Reminded though that shares knock up against zone pop issues.

OK, so on to making it easy to compare different bits of the code. Let's see.

It might not be functionalness so much as clearly breaking down what I'm doing with each dataset.

I do need to redo the scots data for the Europe error, but let's not yet (to not change stuff from before...)

..

Another thorough data comparison. Some small differences that I can't imagine would make the difference. But at least I can use that to compare like with like.

..

And on with functioning up. I'd like to be able to output regressions by just passing in whatever the factor is - pooling each as its own separate run. Err. If that's worth the time. Let's look - first-up, want to repeat Scots regressions.

OK, some while trying to fix something that was working, a good stage to stop for a bit.

..

Looking good. Might also have picked up something that was buggering with it before, but we'll see. It would seem like a bizarre coincidence: labels repeat just fitting... ah wait, actually, it might have done, as long as it was divisible by three (pasting in for all census years) ... but then it would have been fine. So that's probably not it!

But we do have it nicely functional - for this particular task. Suspect it's currently a little brittle.

### Setting up different regressions now it's getting towards functional

Just need a table to break down what I'm aiming for in total. There's a few.

* GB-version of Scots data compared to original Scots data
  + Zone breakdown:
    1. All zones
    2. Original Scots urban
    3. New pop per km2
  + CoB breakdown:
    1. Need to compare exactly what's in old and new Scots. Including exact matches vs "same name but was added up differently". If it wasn't added up differently, we need to check what's happened. (Note orig Scotland's got its Europe other wrong, or one of those cats.) So compare:
    2. Pick out only exact-same countries for each. Difference there?
    3. Compare different CoBs vs rich/poor

Then see how GB looks once we're sure what differences that all makes. For each of those I can get a different bunch of subsets for the new function. Then I want a function for actually outputting regression to text file.

Remind me how those regressions get output... OK, so I've got it working for single output and by factor in a single function. Nice. Now I need to set up listing all the outputs for the above and comparing.

### The actual regressions

From the list above. In theory all I need to do is create a bunch of zone and CoB subs to run through then output. In theory it shouldn't take long.

I should make sure I know what the exact differences are between scots and GB categories. There were some different binnings. Not sure I recorded what all those differences were.

* Scotland all zones split rich poor:
  + All CoBs (different for both)
  + Same CoBs
  + Same but only exact matches

Getting somewhere: those look the same. As they should given the data's identical. But this is necessary for tracking down why there are differences. Oh, I should say:

* This is only for those CoBs that are:
  + From Geoff's subset of Scots countries
  + That are common to both old/new data

So the next thing is to compare for the different data in each, see where the weirdness comes in. (If anywhere.)

### Tabling up

To make it easy to compare. Just the two main geog coeffs... though should I include r2s? Probably. I could automate that as well with some more faff, but... well, tricky in that I'm not sure I remember what zones I was using in the original.

OK, I think I need to compare the w91q weights matrices between the two. They're too different. I did check both of them at various points in QGIS so I can't imagine this is it, but... err, could it be accounted for by the other regressors? Dunno. Seems too extreme.

..

Well. That's interesting. Just re-run on the actual original eviews file and it's the same result. So best go back and look at what I ran before to work out how the bejeesus it came out with such different numbers.

Let's just check it's not the additional regressors.

Ah: **it is.** So that means there's something else impacting on it - this is for **rich/all zones**. So I think it means population density is a more important predictor. It's nowhere near so important for urban only. Well, now I have a clue what's going on then. Nice.

I think I need to check whether **individual countries** differ in any important ways. Or not really: I already know which match. I just need to check that the **same** COBs are producing the same results.

# Towardz an actual paper? Maybe? Maybe.

## So then.

Indeed. Some chats tomorrow, briefly. Something copied over. I need to not be arsing about too much - there's a simple job to be done writing up the method. What I've currently chucked into the shared folder paper is brief, there's more to say and better perhaps.

Oh, the place this is now being written:

**C:\Users\SMI2\Dropbox\Paper\_2 Scotland spatial persistence**

So the first thing might be reading the little bit I've actually got already...

### Now then

So I just want to write here in my own words what the method was for putting the data together. Reading it, it's probably mostly there. I just need to be a little more clear on the essentials. Somehow. Can but try.

..

Well I'm not sure how easy it is without knowing what the rest of the paper's about. So I might just look at employment again. I did actually send something ...

### Random writing

Trying to get to it. Thus.

...

It's a quirk in the 1991 Census structure that dictates the choices made about harmonising the data we use. I will explain it thus: the 1991 Census has two sets of tables, each of differing levels of detail and anonymisation. 1991 'Small area statistics' (SAS) tables, as the name suggests, give data at the smallest geographical zone available in that decade. For Scotland, these are ... output areas I think? However, because they are so small, disclosure risk is higher and so - for the country of birth SAS table - the number of categories is quite small, only x. 1991 'Local Base Statistics' (LBS) tables, on the other hand, are at a larger geographical scale and contain a more detailed breakdown for country of birth.

The 1991 LBS tables, however, have another issue: disclosure conditions mean that some zones with small population or housing counts are set to zero. The population counts in these zero-set zones are then assigned to neighbouring zones, in order to keep overall counts correct.

..

Which is a good start. But let's read back through the actual thing again and work out some bullet points of things the metho write-up needs, how it ties to what we're actually doing etc.

### Bullet point this shizzle up

Yeah, let's try this bullet point thing because trying to explain it with the stuff I've got is proving to be bizarrely annoying. Thus, the things I need to say then to get it across in some sensible way.

* We're using a 3 census span of data and a 5 census, looking at CoB and employment levels in particular. The 5 census picture for employment covers the most recent tumultuous history of change. (Some stats on that wouldn't go amiss while I'm looking for info.)
* The creation of a unique geography: why? Why good? Pivots on 1991 LBS/SAS census table structure and creating a new geography to circumvent zero-count zones. This then becomes the natural geography.
* Geography again: show something of the relationship between larger geography and aggregating from smaller values, for the other censuses. We know by using the 1991 LBS / agg approach that it's as accurate as the Census is.
  + Prob worth mentioning that for pop base we're just using totals from CoB. Anything else wouldn't make very much sense, though it'd be worth dropping in some refs. My point as always: little point in spurious accuracy.

..

Then there's writing up the other vars a little bit that Geoff uses. Dropping the full equation in and writing up. A few more words there.

* Queen contiguity, proportion other non-me CoBs.

### Fixing the Scots CoB thing

There was one category I got wrong (turned up when doing England). A nice little discrete job to fix that, if I can recall what it was and how to go through this entire process again even though I thought perhaps finally I'd got it done. So first-up, find out what category it was I got wrong. Here's what it is:

"Oh dammit: 2001 - non-EU countries in Western Europe" ...

And again, I can't remember if I harmonised the categories or the geographies first. I think it was the categories, right? Yeah. So. Ah: it'll probably need re-doing in both 3 and 5 census versions. Which is marvelous.

Let's start with 3.

Actually, I need to look at the categories. It might require something more than just one re-assign.

..

OK, so I've just worked out **I've already updated the 2001 scots datasheet.** Must have needed to do that for the GB data. Just so I know where all that was:

**CoB\_columnNameUpdate.R**

So I can now update the colourcoding thing to see.

**UPDATED\_APR17\_ColourCoding\_1971to2011\_CountryOfBirthCategories\_Scotland\_noRegionHeadings\_allCols\_No91SAS.xlsx**

..

So now I can look at harmonising again. Starting with 3 census.

First part done. I think there was an error in double-counting china I corrected too. All very familiar-feeling from doing GB/Scots.

So that now leaves going on to geog reassigns (and direct 91LBS). Before doing the same for 5 census. Wherever the hell that is.

..

Enough for one day. Next, need to go that 3 census geog stitching. Then find out about the 5 census stuff.

..

Back for geog stitching. Let's see! 91 is different, as always... OK, that was fairly quick. 5 census...?

**CoB\_stitchingCategories.R**

..

In theory, that's all the CoBs now turned into the raw shapefiles for 3 and 5 census. Because I'm sure there's been no mistakes re-running all of that. Hahahaha. Now to make the CSVs from that. Which is in...?

Err. OK, so done that. (miscDataReshaping). May have accidentally got some order wrong for employment. Oh - so I now need to make the single files for CoB?

And then? Well, let's look at the wotsits. Or, call it a day. Seem to have done enough.

## Employment again

Something to try and get done for tomorrow, as well as try to get my head back into it: what patterns, if any, can we pick up for Scottish employment changes? I already did some work on this so need to go back and look. I thought it would be difficult to pick out any patterns, but if I could find some way to show them, that'd be a start...

Oh, and if there's a CoB / employment level cross-tab at the aggregate level, that would be useful. What did I do before then?

### Correlation of emp to share

Not letting me paste it. I'll go back to it. So this is right at the start of **regressionsOnEviewsReady.R**. And shows that, likely, TTWAs are a good overall geographical level for thinking about employment per CoB. Possibly.

Though not sure where that gets us. Need to think what the point was.

..

But that's just 91 employment vs 11 CoBs. And what else was there?

..

Just realising as well I need to get the data back into the eviewsready shape it was in. Having redone the CoBs. The rest should be OK, right? I don't know. This could take a little while to work out. Christ there's a lot here and it's messy.

..

In conclusion, no, there's not much done on the employment numbers. I thought I made some maps... oh, well, the original data output as shapefiles so that might be what I'm thinking of. I want something else.

I might as well carry this on in regressionOnEviewsReady, though I'll need to go through and fix CoBs again at some point. Perhaps not today.

Err. I did all the econ-active bar chart breakdown stuff last summer didn't I? So what am I after for this? Something different, change in employment levels.

OK so let's try and think about what I might be able to get out. I thought I had already done this, but I guess I need to start by re-doing the basic employment %s and then changes. Thus.

The basic... well, I'm putting them all here:

**R\_outputs\5censusEmploymentChanges**

And now I wanna know about average **change** and where different zones come into that. I could do with knowing about change clusters, for which I could do with contiguity averages. (For which I could do with a function for that.)

Or could just look on a map! Clusters should be obvious if I pick out higher values. Quicker / easier. So just initially looking at decadal changes. There's probably a quick way to do this. In fact, I've done it. How did it work and I wonder where?

Well, just did it manually, is fine. (For a given meaning of 'manually'.) Some striking diffs. So all I need to do, really, is map that shizzle.

..

I can also look at places that had drops two decades in a row. I think there are some. But for now...

And! I want to know if zones that dropped massively - were they very high employment to start with? I think I want all five employment stats visible in the map.

..

Nice graphic idea:

* Graph 5 census EA changes with line for one zone. Strong.
* Get contiguity, graph those, but fainter.
* Get next neighbours, repeat, fainter.

Get to see something about both localised changes and change over time. And I have a bunch of code for that somewhere.

..

Well I've got something now. I could do with something more for tomorrow to actually tell them. So what's a reasonable / quick way to approach this? Errrrr. I'll attempt to do a little of this tonight I guess. Then there's a little more time in the morning, though it'd be good to send it before the meeting really.

Might be worth just starting to write up summat... And mulling how it connects to CoB I guess. Particularly rich/poor patterns. Though just identifying some clusters is probably a good enough start, and I think we have that.

Just mulling what it means for employment to drop. Over ten years, how much of that is a change in who lives there rather than anything else? Hrrrm. Gwilym thinking about the same things for centralisation of poverty stuff.

## Paper bits

Assembling what Geoff has got and what I've got. Did Gwilym write something? Don't think so. Sticking it in one place. Err. OK, that's done! Err.

..

## Re-doing Scots from sorta-scratch

Being done in **regressions2.R** in an attempt to bring together a few bits I've got dotted around.

Three relatively simple things to get done from Gwilym's notes. Two weeks. A few other things to get done, but...

* Send latest model results on 1991 to 2011
* Regress at postcode sector level: house prices 1991 = f(Census variables, physical attributes), then predict for 1981 using 1981 Census values, and for 1971
* Run model on 5 Census years for Scotland, and for 4 Scottish cities.

..

I also want to get some writing done on it. And there was something else. Err. Anyway, let's start with the essentials. The first thing I need to do:

* Table out what all the variables are again. Make sure I have them. Work out what's needed for the 5 census. (Which should be all the same variables as for 3, plus this house price thingyo.)

So I've done that before somewhere. Migshare minus own is the annoying one. It might be easier to look at the doc rather than trawl back through. So here's what I think they need to be. Actually, here they are from the eViewsReady sheet. OK:

* **CoB shares: XiJs** for each decade (actually only got vars for 91 and 11 currently. Might stick to that?)
* **popsh91: each zone's share of population**. Err. Note I'd need to recalculate that for the zone selections as I'm now doing for COB shares. (And if we do the remove-furthest-zone thing).
* **ph91: house prices in 1991**. So note that means we might only need them for one decade for the 5 census data.
* **pop91**: **actual zone population**
* **hs91: count of dwellings in 1991.**
* **pophs91: people per dwelling / crowding**
* **migsh91:** just failed to work out exactly what this is. It's summing to some stupid number per CoB (2529.043) ... Oh I'll work that out later. It's the "stock of migrants in the area from other countries" and I think it's just the sums of the shares. Can check in the code (and ask cos now I think about it, that's odd.) Includes **migshare-minus-own**.
* **Econ active** and a bunch of other vars suggesting that EA at TTWA level is probably the best at predicting.
* And house prices and crowding are logged.

..

Just remind me of some of the code then. Putting it all together into the sheet. Except... wasn't I also re-creating some of them each time? So some sensible code for that? Oh yeah of course.

* There's now a function for making the CoB and weights data. **Function\_CoBRegressionFunctions.R** currently used in **englandWalesDataWrangling.R**
  + That needs the migshare bits added.
* Then can perhaps add a further function (taking that result) for adding the extra bits.
* Ah note: this function is **taking in the original CoB shapefiles** (or a list of them). Not the CSVs put into one sheet I did before. Urgh.

..

* And a quick look through eViewsReady...? (Bearing in mind I'm aiming for 3 + 5 census sheets. And ideally separating this out as much as I can.)
  + It's taking in data with rows being zone and census, long. (So zones repeated 3 times for 3 census.) That was done in **miscDataReshaping** I think. So that's neat. (Urgh again.)
  + Mostly I probably just need to re-piece it back together again. Esp with the new data. Oh, but at least it comes straight from the shapefiles, no middle-thing. I should do that with the rest of the data too.
* The total population column will need doing in the CoB function also. Probably.

..

Small amount of coding done: totalPop now in the function (so total pop proportions correct... those are used, actually, aren't they?) It needs generalising to more censuses and currently only keeps one decade (1991) for weights. Should probably add the lot.

Oh, and mig-minus-own. Which for some reason I'm saying is within-zone proportions. Not sure where I got that from.

I'd probably better not start with the generalising code just now, do it in one sitting best.

## GB again: looking at 5 census

Cos there's a presentation coming up and I'd like to be able to say summat about this. What would need to happen? Urrrgh. Let's try and press on with this. So for the GB stuff it's all happening in **englandWalesDataWrangling.R**. So now I'm at the end of that script with **FIVE CENSUS GB PROCESSING~~~~**.

So the CoB cats need harmonising before the geographies. I think I remember that this time. But this involves having to download the data to start with - not sure I've done that. Err.

It will probably involve harmonising England/Wales first then Scotland. Though... categories should be the same for 71 and 81? Possibly? Let's see!

Actually I think they should be the same for all three (you can download all three at the same time... hang on, why didn't I just do that? Try again!)

Actually, turns out can't be bothered to do this right now. But the thing:

* Just download 71/81 for all vars - code already exists to wrangle it fine diddly fine. (Though I don't have anything but CoB for 91 to 11 do I...? Oh God...)

..

Right, so downloading again. For all GB 71 81.

Oh I seem to have already done it for 71.

* 81 I can actually select a "total" column so don't need to sum M/F. Was that the case before? Anyhoo.

..

71 and 81 CoBs for GB done. Next is...? Harmonising them. Time to stare at it. Did I have a different list somewhere for the slightly different GB data? Oh wait, was it printouts and scribbling? This'll be fun!

So yeah, there's no way around some staring at it to remind myself what I did. The previous scots/GB harmonising may not fit here. Err. Let's check what I said above.

..

Err: looking back at how much of a huge pain it was, might I suggest thinking about merging with the 91-11 GB cats **I've already done**? Let's look and see if that's in any way doable. (And what I did before...)

And read back what I did before to check what I know about the categories. Oo exciting.

Well, I'm reminded that it's bleedin' messy.

Actually, it's possibly quite easy. And possibly I really need to go back and remove some of the others (Africa in particular) from the original scots 5 census results. I don't think the colonial split maps neatly onto anything. Could be wrong - probably needs going through again. Probably.

But let's state the rule so I can defend it:

* If a category cannot be exactly lined up across decades, it has to go into "other". Otherwise you don't know that exact % splits between the categories.

Which then only leaves me with a small number. Let's see.

## GB: apparently I'm presenting

In a little less than two weeks. I should probably do something! It won't be much, mind. But. Let's see. At the moment I'm just staring at maps trying to figure out what might be interesting about the overall pattern.

..

This sentence just occurs to me:

* Here's the difference between looking at in-zone percent and across-GB percent.
  + The latter lets you ask: how has the distribution of people from x country changed over time? Are they in the same places, different places, is the pattern changing?
  + The former lets you ask questions about changes in number, geographically.

Or something similar to that. Oh and: if I can do 5 census, I can say something about how non-UK has changed. So that's good.

### GB 5 census

OK, it turns out I do want to do this. But for this event I think I'll just stick to the main categories that I can get, not worry about the complicated ones.

So where had I got to with that? Oh God this is always so murky! Staring at code time... Oh and I'd written about it above.

So I'm proposing to use the previous 3 census GB as the basis for harmonising to 5 census. Which I'd started to look at in **englandWalesDataWrangling/Checks for 5-census GB COB harmonisation ...**

...

Well that's taken half a day just to get the COBs recoded. Always a massive faff. Now just for 71 and 81 to be zone-aggregated. Reminding myself how. The CSVs I've just made for 91 to 11 will be fine though they'll need attaching to a shapefile, if I'm to do any of that stuff. (Which I may...?)

Ah. Small thing: **I have to assemble all the GB shapefiles for 71/81**. Err. So I've got Scotland. Actually, do I have Eng/Wales for something else?

And I have to do all the dissolved zone ID shizzle as well. Ho hum. Let's download em first. I seem to have 1981 for Eng/Wales. Need to get 71.

Let's look at them... do I recall there being problems? Err. Let's see! If there are, I think I ... let's just see...

* 71 EDs for each country tesselate fine. I might merge them in QGIS before ID-merging to get some all done in one.
* 81 EDs tesselate fine too. So! Merge in QGIS then ID dissolve.

Now for dissolves, saved into the right place. Note: I'm going to need to let these sync with the work machine at some point, they're still on the C drive only connecting via resilio.

..

OK, geographies prepped. Phew. Now I can... err, how does it work again? Let's look!

Aaand another job I forgot: **intersect shapefile for PCSs and 71/81 EDs.** So need two of them. Onward...

Aah, for some reason 71 EDs dissolved by IDs don't have Scotland. What happened there?

* The problem's not in the QGIS merge, the data's all there. Something happens during the ID dissolve. Which is going to be a pain. Can I get QGIS to do it for me instead?
* Ah, just thought, maybe Scotland doesn't **have** a zone code... let's see what QGIS does... Ah yes, I think that's it. Let's look again.
* Yeah, Scot just has ZONE\_CODE in capitals, needs to be lower case.
* Oh fucking hell, it's going to be easier to load into R and change...

Ended up just doing the shapefile merge in R, which I should have done to begin with. OK, now then. Oh – dissolve by ID again. Which should work this time, right? OK, that's going to take a little while to run...

Ah in the meantime, the 81 intersect... And now 71 dissolved by ID ready. So in theory, nearly there with the geog data.

Ah. Did the wrong 81 one. Didn't do the ID dissolved one. Right. Running again. Fine. Man, I need to go through and write down all the stages here. Well, in theory. That will likely never happen of course.

..

OK, so while they're running, some thoughts on what I'd like out of this five census data

* A long view of how non-UK populations have changed (or haven't since 1971).
  + So the non-UK breakdown works fine as a starting point for talking about that. The story should start there and build up towards any extra detail the 3 census can provide.
  + As mentioned above, there's the absolute numbers and how they've changed. But the pattern is also key here: how that has shifted over those 5 decades.
* I'm just mulling... ah yes, this is what I thought of before, wasn't it? You can make a graph that traces one zone but lays underneath it various numbers of nearest neighbour values too, over that time, to try and discern what the average kind of pattern is.
  + So I'm then trying to think about how you characterise different types of clustering.

..

So in theory I may now have everything for the intersect.

Hah! We have the data – presuming 91 to 11 are OK. And presuming anything seems foolish, so will need to check. But maybe hopefully. Time for lunch though.

I do probably want to look at maps to start with.

### Looking at the 5 census data

With a mind to getting stuff as quickly as possible into the presentation. I'd like a view of UK/non-UK and how that's changed. And I probably want to be looking at ... hmm, remind me of which is the most useful depending on what I'm after.

So say I want to look at a UK/non-UK change on a map.

Err. So I have some maps! The European one is particularly interesting (if accurate).

Err. Would a simple graphic be the following?

* The old compare-ranks-change thing? So rank of % for zones in 71, then how that's changed for each subsequent decade (or just 71 to 11).

As a corrolary to the maps. That should be easy-ish to do, right? I mean, it'll be a... actually, it will be a fucking mess. Unless I select the top x and see what that does.

Though that won't capture e.g. the European map.

### Europe?

And possibly "how have the nations changed" also. But I think if I wanted to pick one story out to tell, it should be about that, since it's so topical. Just to illustrate what the data's good for and to link together 5 and 3 census.

Oh look, I can edit the style files. I wonder if that works with legends in print composer etc? Editing the style files a nice easy way forward though.

Somewhere in there is the variable it uses too, so could do that? And then programmatically change... this might be a little much for now! Still, having some handy code for setting up my favourite things would be jolly useful.

..

So now let's load some 3 census data to look at. I also want to be seeing... well, let's look first. Did I even make the CSVs for that?

No, so actually I need to repeat the code I just did for 5 census to get it into zone % form. Oh, maybe I have, it's just in an odd place.

..

Err. Got me 3 census data, not really getting anywhere with it. I think picking this story for the presentation is good but will require rather more digging than I've been doing.

Also just trying OS open names to see if that helps me get names on it a little more straightforwardly

In the meantime I should just probably start the presentation huh?

..

### Wibble

Just copied this over from the migration presentation doc – think it's best keeping all of this in one place for now.

..

Got some more data now, starting to actually put the thing together. Err, remind me how much time we have... Session is an 1 hour 20 so 80 minutes. Four presenters... something around 15 minutes would be fine. Which is quite a lot I suppose, given ... Yeah, let's not worry about that and move on!

Just had these thoughts:

* Actually explain and link to the datasources (maybe write that up on the blog). Just so people know where to go if they want any other census stuff – they might not know. So that would include "use NOMIS not infuse, infuse is rubbish. Scots data is awesome. Syncing all of these is a pain..."
* And! There should be a graph over all the censuses for Europe to show total numbers but with 1991-2011 ones breaking it down by the extra categories we've got.
  + Hopefully those match! Though it may be tricky to get those to look neat, given how ggplot deals with zero points. Can see though. Stacked bar probably best...?

### Europe graph

So doing what I just said as a start. I can then mull how it might be broken down to get something more interesting.

Oh good: the data's fucked. Pretty sure it can't be right but will have to dig through and check it all...

Ah, that was quick: wrong auto-loading of filenames. Let's see if that helps... Yes, that does!

I should probably start putting slides together now, though I'm struggling to stay awake.

### Non-UK for all of GB

Let's make another graph to cheer myself up. Though it won't cheer me up if it takes all day! The plan:

* Boxplot of non-UK % for each census year
* And each-data-point alpha'd for the same
* And then for each of the nine (eight actually, no NI) regions, shifted. Whatever the ggplot word for that is.

Well that took seconds. Alpha thing not good here. Trying some other things. For a start, I'd like to look at where non-UK > 10%. All outliers on the boxplot.

..

Interesting tidbit:

* Scots 2011: a lot of the rural areas are > 5% non-UK born. It really stands out from the rest of the country. I might look at the original Scots data to see where they're coming from.

So I could really do with pulling some information out. What do I want to pull out?

A modal map would be nice. So:

* For those zones in a certain bracket – and actually the 5 to 10% (for non-UK) one could be interesting, given that the focus is usually on the highest: what's the modal CoB in each zone excluding UK and Ire Rep?
  + And this should include the overall count for those zones on the map, so in each of the brackets, what's the non-EU or non-UK count, and what's modal non-UK/non-EU group in each?
  + The story I'm telling there: rather than always concentrating on the high-% areas where all the focus is, all the policy focus, look at how the entire picture changed over this time. Much more widespread, much more of an entire change.

### Doing that modal shit

To help build up this story. Which I can do for 91 to 11 and see how it changes. Less interesting for 71 but it might pick up summat interesting. Let's see, what does this need?

This is really interesting. And actually it will be worth doing it for 5 census – i.e. to pick up on differences between Europe and other places. Some striking patterns (with disgusting colours! Might be worth getting better colours and looking at a smaller number.)

..

And! Remind me what the other thing was. Or what I'm doing or where I am. Err.

### Modal for 5 census

OK so that's sort-of interesting. Err. Now where to go with it? Sort of need to wrap up – and **not** do extra 2011 work on the modal thing. It might be worth pulling out what I've got given how long that's going to take. Then finishing off the explanatory shizzle. Thusly.

### Getting the story

OK so there isn't much time. Not quite sure how little yet, waiting on Gwilym. Possibly very little.

So I need to knock the slides together from what I've got. Let's just get the two overview ones from Rstudio done. Or three. OK. And then... maps...? Let's just list what we've now got and think what to do with it:

* Non-UK: as with Europe, overall, this is a story about a change in the South, perhaps, overall.
* Europe: how the smaller numbers have expanded across the UK, that most places now have Europeanness in them. (If I have time it would be good to repeat modal just for top-European-country.)
* England: would be good if I can drop it in, just to make the point about where you draw boundaries.
* Modals are all interesting, for different reasons. Need carefully explaining. More important than England.

So there's stories about:

* The UK as a whole.
* The south, I think.
* London probably worth mentioning though it's what everyone pays attention to.

OK, let's just make some maps... right ee ho, that's a bunch of lovely GIFs. Have to think if there's anything else to drop in. The modal shizzle, yup. I might still (later) try and get some more detail on the modal shizzle ... except that has nothing to do with census-linking, so mebbe not.

So let's see what's interesting about the modal.

Oops, did everything in WGS projection. Oh well. Changed back now I have better underlay map.

That might be enough maps now. Need to think about getting this finished.

OK, I'm arsing about with other data. Stop it! Enough for now. Write the fucking thing up, move on.

### Finishing the pressie

Oo, might try and slip the English in there. Err. Fnarr?

..

So I'm eshewing (mostly) the full written story. I am, however, learning from before and putting all key phrases into the slides. Which I now need to do for the maps. So let's see if it works.

Err. Nearly eight o clock, don't really want to still be here at 10. And need to practice. At home? Here?

See this?

"... but the more widespread change from red (less than 1%) to orange is the bigger story perhaps."

That is, London vs South. I'd like to know the total numbers across the South (though it's ... yeah, actually, it's a dumb comparison.)

## Back to the Scots paper

### And firstly

Is working out what was on my list of things to do. And getting the first of them done, you know, the important bits. Err. So what's on that list? Checking the emails...

OK so just picking out a couple of things to get done. I already have the code for this, it's just re-running it. Only one category has changed. But there were a bunch of different breakdowns remember?

Oh and I made a better function for it. How exciting. So all I'm doing to start with:

* Re-running regressions for Scotland. For four cities. For all. Probably for Urban. Four cities...? [Glasgow Edinburgh Aberdeen Dundee](https://en.wikipedia.org/wiki/List_of_towns_and_cities_in_Scotland_by_population).
* Bit of writing: Fotheringham, various other things on migration. Stuff on stitching.

..

So finding all that code should be easy enough huh? Err, wait. We also need various other variables in there. Which of course I've forgotten. Let's just do a stare at the code to remind myself.

OK, so first up, where's the extra category gone I spent so long putting in? Is it in fact already there? Err.

Oh wait - it wasn't a new category, it was a change in the count, wasn't it? Err again... So it was just a difference for one decade in 2001? Really? Yes. Search:

**"Oh dammit: 2001 - non-EU countries in Western Europe"**

...

So there we go. Can now carry on with just getting started. Actually I'd done some of this above... [Redoing Scots from Sorta Scratch](#_Re-doing_Scots_from).

OK, let's do the easy thing. Which means not worrying about migshare minus own for now, just drop the rest in. And mark out urban etc - which we should already have.

I can't remember if there was any good reason to have the extra zone pop total column at the end. Is it there for working out migshare? Possibly. Maybe drop for now...? (It was clearly some work to get in there!)

..

OK so having done all the faff I can, can I get on and do this?

### Migshare

Just nabbing code from eViewsReady. Was just trying to find the equation. Hmmph.

Sort of getting there. Think I might only get the things re-run at this point. This was a stupid amount of work to end up where we were months ago.

OK Kittens, I need to go to bed. No wait, the other one. Err. Home.

### OK so knocking together the actual regressions

Or nearly. I just need to drop in density, don't I? Though proper density would be something different which is why I left it out but let's do the basics. So it's houses per acre.

..

Sent a doc. I should probably get a sense of where other things are generally. The list of jobs again (not sure G sent me the latest one but here's this one):

...

---------

Actions:

---------

\*\* Geoff

    -- send Dan a list of the Census variables he and Andy used to predict early HP averages for Census years.

    -- draft this economic literature review

\*\* gp

-- reply to Igor re presentation (Dan to provide slides from his talk on 15th June)

-- send Geoff RCI paper

-- once Geoff has sent his draft literture review, write paragraph on networks literature -- explain the reason for spatial persistence

\*\* Dan Step #1:

-- send latest model resutls on 1991 to 2011

-- regress at postcode sector level:  house prices 1991 = f(Census variables, physical attributes), then predict for 1981 using 1981 Census values, and for 1971.

-- run model on 5 Census years for Scotland, and for 4 Scottish cities.

\*\* Dan  Step #2:

-- Economic Shocks:

-- how does the spatial persistence parameters vary across areas

-- compare rich migrants and poor migrants

-- sensitivity analysis on city boudaries (see R code used by Meng Le to sequentially drop the most peripheral aerial unit)

-- compare with descriptive analysis that helps us tell background story.

\*\* Dan   Step #3:

-- estimate measures of segregation for each the sample years -- for the 4 main cities in Scotland

     -- liaise with Meng Le about applying R code includes Index of Dissimilarity, RCI, local RCI, and possibly now Moran's I based on distance.

     -- 3 Census periods, and 5 Census period (CoB category different)

    -- run with and without Irish?

     -- non UK born vs uk born (exclude the Irish)

\*\* Dan   Step #4:

 -- literature on harmonising geographies

...

So. What's next for me? Well, probably the writing to start with huh? Well I could probably do with reading what's actually there too. To make some suggestions on what we're doing next. Err. Hey, let's do that now!

### List of things

Post-meeting. Basics are more regressions, though I now realise I could do with employment for the 5 censuses (which I... oh wait, no, I do - it's only for GB that I don't.) So:

* 3 census 4 cities rich/poor (just urban? Probably)
* 5 census samey wamey. We don't have a bunch of the housing things, so the rest (and maybe just 'number of people per acre' as a rough thingyo. Err. No point?)

And then the write-up, which I need to do by Friday. So my week is planned out.

I also then want to dig a little more into the meanings. But not until the rest is done.

### Extra regressions

So let's just knock out the cities.

Ah, update: I've re-run to get correct proportions for each city and they're quite different values. Gonna A/B them in a table.

OK, I'll send in the morning. Getting a bit spangly, need to check it makes sense.

### Unit roo!

This is utterly sublime.

<https://stats.stackexchange.com/questions/29121/intuitive-explanation-of-unit-root>

Ah and a useful idea here too. Being able to shift to any point in the time series through the lag operator's exponential (cos it's defined that way as a multiplier, so...)

<https://en.wikipedia.org/wiki/Lag_operator>

### Weights for isolated zones

Cos when I get urban zones for cities, some of the zones end up isolated islands. And if we just use those for the **weights**, we then end up with zero weights as the other zones have been removed.

So code needs tweaking - or at least having the option - to have weights for all surrounding zones that we're able to use.

Oh actually, tell you what - I'll leave that as an open question, as there are actually reasons not to do that. (E.g. weights for zones in cities should just include those zones.)

# Moar scots: five census regressions

## What do we already have?

I probably ran this somewhere already, right? But now I'm doing it again, so the first thing I'll be needing to do is generalise some of the functions a little to make it work for different numbers of censuses etc. Orrrr just hack the fucker for 5 census for now, might be the better option. (Given we only need two decades, actually.)

That said, I might also run 91 on 71, to see how it compares. But 11/71 to start with.

Err. So a hacky function first? Oh wait, did we also already make the 5 census sheet? Let's us begin the vaguery.

Working in **Function\_CoBregression Functions.R -- CoBtoRegressionReadyData**

Generalising.

OK, got some of it generalised. Some other things that need fixing in the function:

* Migshareminusown needs a distinct working out for what's UK, for each. Might just hard-code that in (other option is passing in indices for UK columns... that might be better).
* That section is also using specific column numbers. Newp!
* Both migshare and weights matrices are using 1991. Need an option to say which decades we want.

OK! Now let's see about getting the 5 census into there. Errrr. OK, getting there. Just got to connect a few things... and arse, am I ever missing some variables. Oh well! Um.

Getting there. Doing it in Rmd. Though I've just realised if I want to do the cities, I'll need to re-run as I was doing before, on the subsets. Looking quite different - but, hey, it's a start innit? I can run on 91 too. Actually, let's do that immediately.

OK, well, there we go. Err, I should do 11/91 as well, to see how different it is from what we have from the other numbers. (Ideally shouldn't be huh?)

And cities! So that actually requires running the whole thing to get a different subset out, same as we did before. Sososo.

..

I want to just try running that all again with a slightly smaller list of CoBs for splitting rich/poor more clearly. Let's look.

Just doing the different subsetting of CoBs

## Writinz!

5 census stuff all looking highly suspicious. Need to do some careful checks.

..

Before doing that, just mulling the unit roo(!) stuff again. Some things falling into place. Sort of. With some bafflement over some of the incredibly basic things.

First-up: it's the same quantity over time. That's the basics of it. That's why it being <>1 means it's either growing or moving to some stable value - I mean, in theory. I don't think that makes the blindest bit of sense for what we're talking about here. But I see the theory.

In fact, something's gone wrong with my brain on this one. I love this feeling of going backwards, it makes me so happy.

I'm struggling with the basics of y = ax + by. That y is a sum of these two terms. This is dumb, this is not something I should be having a problem with.

<http://people.stern.nyu.edu/wgreene/Statistics/MultipleRegressionBasicsCollection.pdf>

That's good. And stopping again because WRITINZ! Let's just look at it if nothing else. I have a churn so this shouldn't be too painful (he says)

Right so, did I actually get through the doc? Yeah, think so. Some leads to follow but this is it.

..

Gonna stick the notes in with what G and G have already done, read their bits, then probably attempt sleep. Though the tiredness seems to have dissipated. So this is a copy so it's not saving in the shared folder the whole time:

**Dropbox\SheffieldMethodsInstitute\Migration\Paper1literature review gm gp 3july17.docx**

..

Noting the themes that they're writing about for combining with mine:

* Intl migrants cluster. There are some basic reasons why.
* Clusters can be disrupted by shocks. Mobility is a function of scale – people move more locally.
* Paper innovation: consistent geog scale 71 to 11.
* "The paper attempts to explain changes in the local distributions of each group across the censuses, **including changes in segregation**." Are we doing segregation in this paper?
* "there is evidence (reference) that incoming migrants displace domestic residents to other locations; furthermore, poor migrants have lower internal mobility than either domestic residents or richer migrants and the combination increases segregation"
  + Some of that fits with stuff I've got.
* "We are particularly interested in identifying areas that have undergone significant change." OK then.
* "We examine the extent to which the parameter differs between country of birth groups, between rich and poor migrants and between groups that have large rather than small representations."
* "We concentrate particularly on changes to industrial structure, since significant changes in the Scottish economy were taking place over this period." OK then.

Now onto the lit review bit where much of mine needs to go.

* Starts on antipathy to migrants. Not entirely sure I like that.
* What do those 1st para refs say about impacts on wages/rents/housing…?
* "Research on the effects of migration on housing markets, which is of particular concern here because it relates to migration location dynamics ... " so it's about housing as well?
* Headship must have specific meaning in this field.
* This is a lot of stuff on home ownership/migrant dynamics. Is the paper about this?
* Good point here though: regional house prices can go up while specific localities with many migrants can see them go down
* Already citing Borjas – I had that one.
* Hatton and Tani, Gordon, Sa: oh, so this is probably where I got that list from.
* Onto how this supports segregation via residential sorting
* "Zorlu and Latten consider ethnic differences in mobility and the location choices of immigrants and domestic residents " … oh, comparing those two. Not read. That's 2009.
* "Non-Western migrants are less likely to choose locations that have high proportions of domestic residents, whereas domestic households are more likely to choose such locations."
  + We should be saying here somewhere – but is this a function of their wealth or age? Cf. counter-urbanisation, mainly wealthier/older/domestic.
* Although this kind of speaks to that a bit: "Ibraimovic and Masiero consider voluntary preferences for segregation as opposed to those imposed by the nature of the housing market. The results reveal some evidence of a preference for colocation amongst the same ethnic group and the avoidance of other ethnic concentrations."
* We're now onto Gwilym's network bit. And racial harrassment. And homophily. Err.
* Polya processes – would it be better just to talk about path dependence?
  + Because it's self-reinforcing, is that why? Yeeaah.

..

And now I have to work out how mine fits in, if anywhere. Err. Note that they've done no citing at all.

..

### The actual writing

Yup, gonna do the writing dump right here and then transfer when done. We'll see about references. Kind of might as well not bother if no-one else is using the same system. So - going through the churn, sentences ahoy. Not tidy. Total brain dump. Sort and quality control later.

...

The literature on modelling internal migration in the UK identifies the key forces at work that determine where people move.

Fotheringham et al (Eng Wales)

Background of counter-urbanisation "involving a net loss of internal migrants from major urban areas to surrounding rural areas." [1644]

Seeking to separate forces common to all migration choices from those specific to different places, age groups and - of course - nationalities of incoming migrants.

"The history of migration modelling is replete with contradictory theory and conflicting empirical evidence" [1640]

Theory identifies opposing forces - Fotheringham et al pick out

The data can adjudicate between these competing theories - though as they poin out, these forces are about the balance of push/pull factors, and so the result could vary from place to place without bringing the theory into question.

[Though if I don’t discuss one of those, won't be much use. Underdetermined theory.]

"well-documented attraction of large urban areas for migrants" [1663] Cities Rees paper.

They identify reasons to expect persistence in both the richest and poorest areas, with quite different causes.

Indicated by a negative relationship between high crime rates and lower out-migration: "Although this is counterintuitive, high crime rates are evidence of general deprivation and individuals in such areas are often less able to migrate." [1664]

Other proxies for deprivation also correlate to low mobility.

As they note: "less able (rather than less willing) to move away from relatively deprived areas." [1670]

Paradoxically this can hold true for newly arrived international migrants connecting with networks already in poorer areas: once there, their position within the internal migration hierarchy is more likely to keep them there. The forces at play that lead them to these places initially reinforce the domestic dynamic.

Ceteris Paribus, "high out-migration rates are associated with areas where household incomes are high, probably because wealthier people are generally more able to migrate." [1665]

Set against this is the tendency for places with higher house prices to see less mobility.

For people from richer countries-of-birth, they would be more likely both to initially arrive in a more dispersed way and, once in the UK, less likely to persist in the same places over time.

..

Framing the issue in terms of stocks and flows useful for explaining what this analysis can discuss. Or something. We can use stock counts for different Censuses to theorise about how flows have changed and how they differ between groups. Most fundamentally, stock changes over time are made up of people arriving and leaving the UK (international migration flows) and those moving *within* the UK (internal migration flows) - with both added to the count of people who have stayed in the same locations between censuses.

That needs something more: a sentence explicitly acknowledging that we haven't separated those out but that it's still saying something useful. The underyling assumption perhaps, that the same forces inform people's decisions to move as their choices to stay put. And again that some of the same forces are at work at the international level as internally, especially considering that arrival into London can be considered a staging post.

I wonder if Geoff's already covered all that? Hmm.

..

Fotheringham et al find that migration churn is surprisingly uncorrelated to nominal employment levels - instead, employment *growth* is important. This makes some intuitive sense: areas with low but growing employment are clearly attracting people.

In- and out-migration also tend to go together: there is a "high turnover effect" [1666] mirror image in more deprived areas, a low turnover effect.

This ties to the london thing: economically 'hot' areas are not only attracting many more people, they are responsible for churning migration movements and thus supplying other, less economically intense, areas further down a hierarchy of attractive locations. [see e.g. Champion et al determinants, interaction effects and how international migration into London might uniquely determine out-flows to other regions xiv]

So the bit that it's going to be tricky to write about: attempting to unpick forces unique to particular countries of birth from the underlying employment and wealth dynamics ... the persistence story is one of path dependency. Geoff's got that. Which is where it differs from ... historical factors that led to particular concentration spots cast a long shadow over future patterns. Feedback in such areas make disentangling cause and effect problematic.

As they say, the data cannot directly distinguish "the desire to move from the ability to move" but logically, wealthier people have a larger range of choices within their chosen cities, and across regions. [sorting already in there?]

Champion et al corroborate this, finding that counter-urbanisation has been underpinned by a supply of people from the least deprived areas - 'the ability to move' again central. [determinants p.968]

..

Ah note: it's Champion et al in Determinants talking of hierarchies. Read 'counterurbanisation cascade".

Let's just try and paint that picture too then (cx 2 London point just above) In contrast, international arrivals may feed into a unique dynamic labelled a 'counterurbanisation cascade' ... actually this isn't just international. Need to read more. Something about de-industrialisation too.

So counter-urbanisation has been balanced, to an extent, by international in-flows, though much more so in the South than in Scotland. [Zat true? Numbers? %?]

[how does all that compare to decentralisation of poverty?]

Other work reinforces urbanness and ruralness (and patterns that have been sorting these) as an age cohort phenonomen. [Hatton immigration and inter-regional mobility p.??]

Relatedly, Finney et al find that, once age and other socio-economic variables are controlled for, differences in ethnic group migration patterns mostly disappear [16].

Differences in country of birth will see some of the same effects

And where am I going with all this? Not sure.

"internal migration is one of the mechanisms through which regional labour markets adjust to immigration shocks." [F342... Hatton?]

## Rotherham bits

So this is for an event next week. Some maps. There's some info at work to get but there are things I can do at home, so let's do them. Writing up here as it involves the same dataset - G wants to see changes over time from the 5 census data.

So the two things to do here:

* Just get the geographies for Rotherham plus Sheffield to see how many there are.
* Put together the employment data for England/Wales ... hmm, I'm having some vague memory I might already have done that. Err. No! Not going to worry about matching up with Scotland just yet.

### Econ active

Getting the data. 2011 the usual pain. So let's collate. Somewhere. Downloads first. Err. I should probably check exactly what I did for Scotland first, really. If I can track it down.

**censusData\_smallToLargeGeogs\_3Census.R**

And where shall I stick this lot? Hum - as a section in: **englandWalesDataWrangling.R**

**~~~~FIVE CENSUS GB EMPLOYMENT/ECON ACTIVE PROCESSING----**

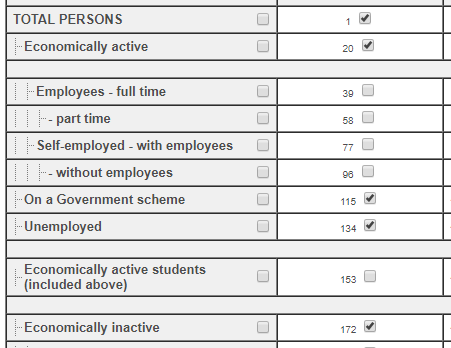
...

All as going along. Think I'm getting the right data but will have to see. And re-do this in the GB small to large code. It's nothing too complex and I have all the intersections now don't I?

..

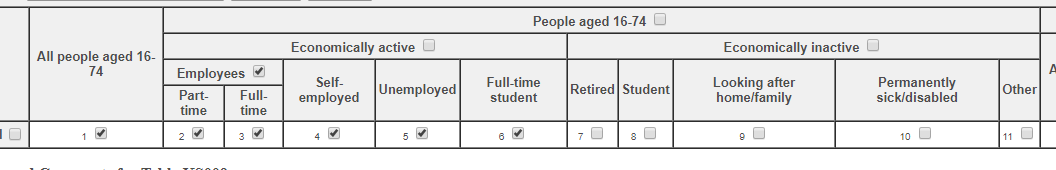
Table notes:

* 1971 EA:
  + Staring at table online to get codes. What do we need?
  + Just gonna go write this in the code comments!
  + Oh except I need to check something on the ages. Back to what I did before... Well, just used the totals. Which actually you need to cos the age breakdowns are only for "econ active" overall. Though that'd be interesting to look at (and how it changes.)
  + Can't find info on whether "sick" is long-term and so outside EA. Hum.
* 1981 EA:
  + total persons in employment (full-time...) = 789
  + total persons not in employment = 859
  + total persons self-employed = 705
  + So issue there of part-timers and general change in defs e.g. age
* 1991 EA (matching wards / pcs):



Which reminds me: the names are -

* Out of "economically active" you get employed/unemployed.
* Anyone not EA doesn't count in that (e.g. students, not seen in the table above, are usu counted as econ **in**active.)
* 2001: needs each country downloading separately. Oh goody. Already have scotland and hopefully it'll join up. Once I work out which columns I'm defo using.
* Update: switched to univariate table to match Scotland. Just 02 and 12 for econ active and unemployed respectively.



Note this appears to include full-time students in the econ active bracket. Don't think other decades did. Did I look at this before? Note: student is also in 'econ inactive' (but not full-time student). Wassagwan?

Then:

* Eng/Wales 2011
  + Probably from nomisweb right? And probably some stitching to be done separately on that. Let's have a look. I suppose I could try infuse...? Ah yes! And got it for all of GB. Nice. Now the stitching then?

..

OK, that's all 5 decades in a form ready for intersecting etc. Now to intersect. Yech. I can mainly just copy CoB code for that, can't I? Then do % after.

Last thing: 1991 direct summing for the agg zones...

Done! Phew. Though unchecked yet. Might just have a look with the next few spare moments.

Note: **one way to deal with changes over time = ranking**. Any subtle changes in definition will not matter if you do that. Its own problems, maybe, but still.

OK, now I have to make maps and graphs and stuff.

..

Ah, another random thought:

* Remember those vectors from GRIT? That was based on sums of flows. Is there a way to do that for (say) ranking of employment levels that could show shifting patterns?
* Tricky. It would need to be something to do with weights changing.

OK, just mapping employment ranks. Another good thing to map would be changes. Err. There's easy code for that too isn't there? Maybe time to leave now though.

### Making the maps and graphs

So let's start a fresh script and get the data loaded. Some in QGIS but let's start with thinking through the graphs.

Err. Where to start? I think I was going to do some zone ranks wasn't I? And at this scale I should be able to do all the maps in R to get things moving a little more quickly.

But yes, to start - zone ranks of relative employment levels, say, to see how things change.

Yeah, no, that completely looks awful. The change is too large. Err. Rank? Newp this is all round a stupid way to approach it! Let's think of something else. Hmm. Time for a walk.

..

OK let's make some maps. So I need some differences between certain decades. So data prep. And - hurm, what CoB to put in there. Non-UK breakdown prob where to start.

..

Right so I've probably got everything I need at this point. I need to tidy graphs and perhaps put ward names in a reasonable order. But it should be enough. Maybe. Yeah, that'll do me. Let's get maps out first.

Enough maps I think. Graphs! And they all need... titles and shit. Let's list what I'm doing so I know:

* Employment:
  + Shef/Roth (and comparison to GB)
  + Roth alone
* CoB non UK
  + Shef/Roth (not comparison!)
  + Roth alone

So just four. OK.

That'll do. That was a day's work. Not a man's job. But a day's work.

..

Little additional job: same scale and more values for the two 71/11 CoB maps. So that'll be saving a scale that covers both after setting the values up. So let's look at those values and pick something useful.

It's actually tricky. Let's see if I can do something acceptable. So the ranges are:

* 1971:
  + 0.9 to 9.4%.
  + Pretty breaks ~2 e.g. 2 to 4%, 4 to 6%
* 2011:
  + 1.8 to 35.9%
  + Breaks ~10 e.g. 5 to 10%, 10 to 15%

..

OK, let's look at the numbers and think about what we can do that covers both. Err. This will work better with equal zone counts / quantiles, the numbers are closer... I just deleted all the working out, really don't need to keep it!

## Back to everything else

### And some random things on the random processes

Reading the lit bit... ah, arse, no, won't bother. Point being: actually making a model that distributes folks in some way – how many different theories on why that happens could the data support? I suspect a lot.

And this also speaks to the issue of explanation that I've just been going over again. And specifically the level of explanation you want. Cf. David's "myth of the return" and some of the deeper reasons migrants have turned up, the role of changes in national policy etc.

Actually, that's a good point: the paper needs to say something about A8 countries and how national policy can be vital. (That really makes that point doesn't it?)

### Extra bits for Geoff

Remind me what it was he wanted... Right, 2011 on each of the prior decades. I might also do the ten year gap thing while I'm at it. But let's just generate the data first. This is where I'll remember just how messy it is.

Looks like it might require me to create a few more pop weight variables – or leave those out since they didn’t seem very much use.

OK, let's try and get this job done. I think I now have all the data I need. Let's set up in Rmd.

### 17.7.17 Geoff Skype

* Did I use TTWA for employment value in previous ones? (And prob not sensible to use TTWA for city-level given there's only one level.)
* For 5 census 71/81/91/01 on 11: split by rich/poor.
* What's going on with the employment numbers? How does that change when rich/poor dealt with?
* Drop to Glasgow/Edinburgh. (Maybe pool later.)
* For me: correlate house prices to employment – they're gonna match aren't they?
* Include house price for 91 onwards.
* Rank unemployment cos some changes in how employment defined
* Interactions? House prices vs employment levels ...
  + E.g. Low house prices / high employment in early censuses where people were living near industries they work for.
* Run 10 year moving lag period for all of the above.

### Idea

* Take any random persistence coefficient.
* Test what range of real-world mixes of movement might fit it. (Randomise, greedy algorithm.)
* How different are they, what's the constriction, if any?
  + This might help think about the ranges of what it means. Which is still bugging the beshizzle out of me.

How would I do that? Java would be the quick way. The point here also is how normalising all zones to 100 masks the underlying patterns. That's the thing I'm trying to get at. You could be doing the same for the contiguity part at the same time, though ... so yeah, that's just about working out what the data-generating process is.

And there's still the issue that the zones are different sizes to begin with.

Thinking of sizes, here's a thought:

* I think it's OK that you regress shares because it's about change. If a big zone gets bigger, that will show up. If they all stay the same proportions, it'll be 1. (I'm unconvinced by "unit root" as how can you have neg/pos feeback if you're taking proportions? Err – well, it at least hints at direction.
  + Actually this is still problematic I think. Large proportion zones, if they're the ones with the larger pops to start with, it's problematic... can come back to that. Didn't I have a thought on this? At the moment I'm struggling to see why in-zone % wouldn't work. I know that if you use across-zone you're comparing 100% in each period and where it's assigned, but... Hmm, yeah, but it's not about whether those zones have grown, it's about the overall spatial pattern for that CoB.
  + So what was my thought? Mix both, was it? In-zone proportions first **then** normalise? Which gets rid of size issue? Yeah, might work.
* But! Also: regressing either of those on e.g. **employment levels**? What could that possibly be telling you? Hang on, I had a mental picture of this a moment ago, let's try and get it back.
  + The larger CoB zones in e.g. 1991: finding a pos relationship there to employment is not going to mean much is it?
  + What **would** mean something: CoB **change** between the two Censuses – which way have proportions gone? How does that relate to employment levels in those zones?
    - Note the issue with employment in small zones really being a proxy for wealth. (I can check this against house cost and some other things.) But it might still show up – especially now we have GB-wide data to check.

But that point about change, and about the weirdness of proportions, I think is probably important.

### Things for Bernard

Let's read the email again. Some of this I can get to him straight away. Other stuff, not so much.

OK, so the first thing is to get everything I already have. Mostly I can use the existing data creation function, just reducing to London zones. So let's list what we have and then what we need to get...

Well, the first thing might be that the data is different for Great Britain as a whole. Can't quite use the same code. So need to get testing that.

For GB, at the minute, there is only CoB and Employment over the five censuses anyway. So that should be a little bit easier to collate. Let's get the first bit done, CoB and weights etc., check it all works OK.

I should really try and get something done on this. Errrr. More tea? And now London, and I'm probably reducing to the LADs in London. Shapefile plz.

This might come in handy:

<https://stackoverflow.com/questions/14208016/find-best-matching-overlapping-polygons-in-r>

But let's just try %in% first.

..

OK, mucho faff: ended up using QGIS intersect to get >50% in LAD zones. Got now – next is running it through to get the numbers. For both 3 and 5 I guess.

Right so... why are there only two decades for the 5 census CoBs in shapefile form? That can't be right, I've... oh no, yes, it can. The CSVs are there, so they existed...

Right, so remembering what I did:

* In a rush for the conference, a reduced number of cats. I used the existing 91 to 11 cats and harmonised those.
* Which is why they're not there. But...
* Short of it is, they all got converted to CSV, so need to put back into a shapefile for the function (or alter the function... let's just check that.)
* Nah, prob need to re-attach to shapefiles. Grr.
* Now commenting out migshareminusown in its entirety – it isn't fucking working and I don't have time to track down what it's doing. Now... chances that weights will work??? Well thank fuck, yes.

OK so where does that leave things? Well I think the code might not (quite) work for three... oh wait, I can fix that.

## Error rate in geog combining

Which for now I probably won't use directly, but worth putting in just to show the issue's been considered and we know where "using just area" falls in that range.

Issue 1: **need error rate for each decade that's not 1991**. Different sizes. Just for Scotland to start with - it'll be different for Eng/Wales (probably worse for Scotland I suspect.)

OK, so we need to look at one intersect and have a think about it. Let's put it in its own script in **census\_dx**.

**error\_in\_intersections.R**

..

OK, where to actually start? Well... loading an example intersection / having a look in QGIS also possibly.

How to check which ones are fully within their parent shape? OK, so area sum is one way - but the issue of slivers comes up. Previous approach to this: ignore. I think that's still probably reasonable as it makes little difference.

But: it might matter if we do want to say x zone is fully within parent. But but: doing so is probably a pain.

So the basic principles, ignoreing slivers for now:

* Say zone x has area split 30/70% between two largers ones.
* We have some value from the data but we don't know where it is.

This seemed clear in my head before. Hmm. So let's assume we know **nothing** about the spread of people within that zone. They could be spread evenly (splitting by area works) but they're probably not. So what other splits might there be?

Well OK, assuming nothing else:

* Zone x, two parts, x(.3) and x(.7) area split.
* The prob of any single individual being in x(.3) is .3 if it's even. We need some... prior (argh) assumption about how that might be different.

Actually, no, that's just it: in this case it's a dead straightforward binomial distribution. For each person, .3/.7. Run several times, get a full distribution.

For larger numbers it's not gonna be binomial.

So the question then: how does.... err. Hmm. No, this approach doesn't work: as the number increases it'll just become a .3/.7 split, obviously.

Ah no, it's this:

* There's a **single** probability of where in those areas each person may be. The area one is .3/.7.
* So prior to any binom (or more: <https://en.wikipedia.org/wiki/Multinomial_distribution>)
  + Set single probability of where housing is. **Then run the person trials.**

### Binomz

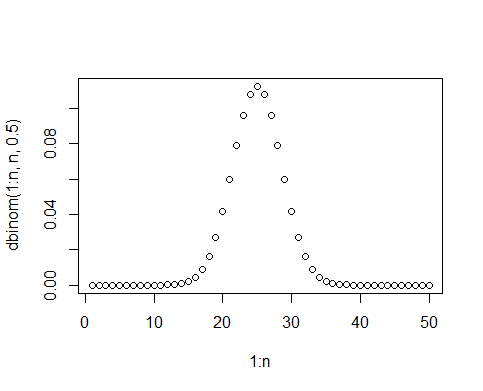
OK, I'm having to go back to basics with binomial shizzle

<https://www.youtube.com/watch?v=qIzC1-9PwQo>

So:

X is just the number of successes in n trials. E.g. how many heads in 100 coin flips. Which with rbinom is easy enough:

rbinom(100, 1, 0.5) %>% table

Right, so: I could re-run that many times to get the distribution, but R gives it to us, doesn't it? That would be... what?

That's using dbinom via <https://www.tutorialspoint.com/r/r_binomial_distribution.htm>

plot(y= dbinom(1:50,50,0.5),x=1:50)

So: first arg is taking the quantiles - i.e. the first part is "probably of 1 successful trial out of 50 trials given 0.5 prob"; "prob of 2 successful ..." up to probability of 50 successful trials.

So just about getting my head around this. Err. So. Err.

OK, come on, I can work this out now.

rbinom(n = 50, size = 1, prob = 0.5)

..

Oh **idiot**. **I'd been confusing myself by sticking it in a table.** This is easy:

rbinom(n = 50, size = 2, prob = 0.5)

Size is the number of trials. Duh. So here, 0, 1 or 2 successful trials per set of trials. So here's a nice distribution direct from the trials:

rbinom(n = 10000, size = 50, prob = 0.5) %>% table %>% barplot



Which I probably won't need anyway: can just get from uniform dist can't I? Err. Onward!

I'm not sure there's any bomb-proof way of doing this. Example to think through:

* 0.5 / 0.5 area split. OK, so then decide the probability is changed to .7 / .3. Why? And what's the extreme limit of that probability? We could assume it's normal(ish) but is there any reason for that?
* Well, better than nothing?

## Dissimilarity index

And possibly some other things, building on Duncan Lee's code. I need to get some Scotland city code - but let's just look through Duncan's code again to make sure I know what I need.

OK, so in the last session 6 PDF: two time periods. I just need to work through this and get things into the same shape. I should probably do both at the same time to check the structure of things is identical.

Basic point - what am I aiming for? Oh, note - I could perhaps do for each CoB or grouping. Well, let's sort the basics first. SO... in Duncan's thing:

* Per zone proportion ... ah ha, immediately that's different data!
* Per zone proportion: so yeah, just raw proportion not % but I guess the math would work out the same.
* So I can start with proportion non-UK born and take it from there.

Which is going to require creating the data from scratch. Hur - for which couple of years? Well if we're doing non-UK only I can use five census data for now.

It occurs to me: **this could be used to compare to story being told by spatial persistence measure** which is more about localised effects. That sounds good to me.

But first, the data!

..

OK, I have something. It might even be something useful. Though... I mean, the underlying black box is, you know. But at least I've got the thing working.

## Restate assumptions

So I don't have a lot of time left. On leave from Thursday. Totally lost track of where things are with this and I imagine G is expecting a full paper to pop out at any point. Though, hum. Anyway let's just check on what the last action points list looked like, if there is one. Well, some bits from emails:

* Concentrate on urban glasgow and edinburgh cos sample sizes. (Fits with dissimilarity stuff)

Also from Geoff, email 10th July:

"I think it is the case that the degree of persistence falls in Glasgow/Edinburgh compared to the previous national results and I’m not sure why this should be the case (also the difference between rich/[poor are not very pronounced). I’m a little worried about whether this is just an artefact of how the data are constructed. Any views?

"The “non-persistence” variables are also interesting and differ a lot between Glasgow and Edinburgh , e.g. house prices."

..

What I sent in those emails:

* 5 Census 2011 ~ each other decades, for all CoB and for Edin / Glasgow. Those need pulling out into something graphical I think.
* I think I also did 3 census previously. Errr. I need to go back an stare at the code huh?

### Getting something written

So can I just use what results I've got? I need to open RStudio and look at it really, huh? All I'm then after is some easy way of showing them all. Err, so let's just stare at RStudio and try remember what was where.

Aaaah I'm just remembering what a fucking hideous tangle the Rmd outputs were. Need to get that better. But using Rmd = good idea. Maybe? Table and graphic at the same time.

Oh and look, there's all the regression functions too. This is ... an unpleasant tangle.

Ah, handily, I have at least noted where most of the hacking is taking place.

**5CensusRegressionsV2.Rmd**: so I don't actually have one for 3 census.

Sooo. Diggy dig time. Starey stare.

..

### Trying quick and dirty

And probably failing. So that means looking at the results we've actually got and concentrating on the contrast between Edinburgh and Glasgow to pull the story out.

So let's look at what data I actually have, try to remind myself what on Earth is going on. Were there some notes from having talked over the data? Probably.

Where was the data, to start with? OK, so:

* I have some detailed 3 census data
* 5 census: all I seem to have is Scots-level, each decade regressed on 2011

So I need to go look at code, right? And pull together some results for urban areas / rich-poor / 3 and 5 census.

Oh wait: PDF was being output from Rmd to root folder. So we do have more. Quick look through.

OK, the four cities haven't been split into rich/poor for the 5 census. Possibly cos that's problematic, can't remember.

..

### Some paper looking-at

I've moved what I think is the latest to here:

**C:\Users\admin\Dropbox\SheffieldMethodsInstitute\Migration\Paper1**

**DO\_update\_compareReview.docx**

From the shared folder:

C:\Users\admin\Dropbox\Paper\_2 Scotland spatial persistence

..

I don't want savings to suddenly being appearing. I think that's the latest but who knows? Kind of completely lost track.

### Into a spreadsheet

Commence swift tedious task... Fuck that, actually. Let's just get on and program it to output the numbers in a better format. And howwww? Well I probably need to look at the actual code I imagine.

Can I at least just look at the code? Err. OK. Did that once. And again? OK, well, that probably kinda maybe makes sense. Though it's a fucking mess. I should be able to pull things out into one place, mebbe? Need to look up that map shizzle again.

But no, first, more time on staring.

..

But not too much staring. The numbers are there - let's just code getting the little bastards out into a useable format. Start with the 5 census stuff. 5censusRegression\_pullNumbersOut.Rmd

Err. Except why do that in Rmd? Let's instead stick it on the end of...

**Regressions2.R**

Taking bits from

**5censusRegressions.Rmd**

..

https://stackoverflow.com/questions/5587676/pull-out-p-values-and-r-squared-from-a-linear-regression

So I should check my understanding of the p values there.

So anyhoo, I can just pull out the coeffs into a dataframe. This gonna be peez.

### Moar collating of numbers and making into pretty graphs

Decide what there is. And look at what's there for 3 census. Right, so 40 minutes timer on in an attempt to actually do something. Here goes.

What does the 3 census shizzle look like?

Also: **sheet5** - **is fucked up I think.** It's got everything repeated ten times. If that was going into the regressions, a good chance it was wrong. Err.

The ones from 2 rich and two poor are correct, I think (**sheet5sub**). 5 census all CoB - something fucked up. I'll come back to that.

..

No, back to to staring at 5 census. Differences again... why? OK, so:

* **Sheet5 has each decade's variables, which is what we want.**
* **But it's got duplicates**
* **And where was sheet5sub made? Different vars in there**.

OK. A mess. I could really, really do with rationalising this so I'm getting exactly what I need to get and don't get in this much of a mess. Or... hmm.

### Checking 5 census function compiling

Something went wrong at a merge stage I think. So I'm going through this function:

**compileSheetDiffZones5census**

With some preformed data. Just getting to CoBtoRegressionReadyData, it's looking like the right numbers: 8220. Which 822 zones \* ten CoBs. So then what goes wrong? Ah, I think I begin to sense what... Let's continue.

OK, the merges now. What goes wrong? Ha: **the very last one - the urban flag.** I didn't select down to unique. I think that's all.

Yup. Hah. In some way I can't figure, the sub one worked. I don't know why.

OK so now I have the right numbers again. Or the right number of numbers, so that's a start.

OK, so where we at? Let me list the variables in sheet5 and think about what's needed for the three census then.

Oh good good: **migshareminusown has vanished**. That one is also, of course, a fucker to work out. Where the shit has it gone?

It's commented out in cobToRegressionReadyData ... for some reason. Well, that'll need redoing! It must have not worked for some reason. Will need to re-instate...

Err. It works. OK then. Carry on.

Maybe it doesn't work for the 3 census...? Guess we'll find out. But 5 is fine. Or was then... maybe subsets don't work. Newp, subsets ran fine.

Um. Well, shall I attempt making some graphs just to see what they look like?

### Getting on with it

I've got the 5 census data together then. Though haven't quite settled on exactly what I'm doing with it. Probably need to pull out all the numbers but need to know what those are. Need 3 census too.

I'd like all of the out. Oh let's just check something...

So thinking of a quick way to get the tables Gwilym's after: R can't split cells but I can do sub-categories, possibly, just by combining columns e.g.:

Glasgow-rich / glasgow-poor.

..

Oh wait. Just looking at it **it's already in the right format**. It would just need spreading and to have better names. OK then. Just need to decide what's going in.

OK, remind me what variables are in the 3 census data:

* Xij / weights / migshareminusown / lph: house price log / hsperacre: density / lpophs: crowding / ea: unemployment
* 5 census is using pop per acre, which isn't really meant to directly be in there? Err, was it in Geoff's? Yup, just PH / DENSITY / CROWD / EMP

Table comparing both. As they stand - probably no point regressing on pop per acre, is the... oh that's in there as a rough take on crowding isn't it? Given I can't get dwellings for 71/81.

|  |  |
| --- | --- |
| **3 census** | **5 census** |
| xij | xij |
| Wij / spatial lag | Wij / spatial lag |
| Migshareminusown | Migshareminusown |
| Employment | Employment |
| Crowding (people per dwelling) | People per acre (no 71/81 dwelling count) |
| House price log | ---- |
| Density (dwellings per acre) | ---- |

..

OK that's useful. Now let's break down the various types of regression I've been running, see which we want to keep. Or... hang on, just a bit of general staring.

Just looking at the code:

* R section: "*Cities: recalculate proportions for each-----*"
  + **First time I'm recalculating proportions on zones subsets.**
  + And I think it shows quite different results for cities. As one might expect. So this is **a different category of thing and I need to be clear when I'm using it.**
  + E.g.: do I use it when doing "national urban"? Dunno.
  + Whenever I **am** using it, need to throw zones through that function. (Or functions given there's a separate one at the moment for 5 census.)

OK, that seems to make some sense, I remember most of what I was doing. Now just to list all the various categories I used. Note - **I'll come back to comparing the recalculating-proportions thing later. For now, use that code.**

* **Source,** **3 vs 5 census:** How to compare properly?
* **CoBs**: rich vs poor.
  + Later: I can also do each CoB separately and Europe vs world etc.
  + 5 census: 2 of each to make sure. And larger group? Well, can throw that in as long as its clear.
* **Places**: four cities all urban, national-all and national-urban.
* **Variables**:stick to whole list for each for now. Some testing to be done there.
* **Time periods**: e.g. for 5 census can run each decade as 10 year steps.

..

### What the numbaz meen

Don't want to spend very long on this but just mulling. I still think there's some basic confusion here. Or it may just be me that confused.

There's just this essential weirdness about what xij11 means when regressed against anything else. Any other regression setup would make sense to me. That is: you are treating it like any other quantity but then dealing with spatial autocorrelation.

Here is different and ... probably time series thinking needed. Certainly. It's not quite panel data though it is values for zones over time.

Yeah, this is quickly getting rabbit in headlights. The only thing I want to know:

* For making the vizs, does it make any sense to have diff coeffs in same bar / polar chart?
* Or: what does it mean to regress xij2011 on e.g. employment 1991?

Let's just stare at that in plot form... OK, yeah, so I remember thinking about this before. It's just weird and wrong.

The idea in theory: larger EA values are determining where migrants go. But you're not comparing to **how migrants numbers have changed**.

Hmm, I'll come back to this. I need a shortcut to the simple answer for the plotting.

..

Staring at this again. Just a little bit then I think I might change the scene / go to office. Thing: what actually are the units we're talking about?

OK, this isn't so hard:

* For a 1% increase in the proportion of migrants there's an x% increase in employment. Or probably vice versa huh?
* And note that it gets tricky when you're comparing different things e.g. xij with migshareminusown. Two different uses of proportion.
* So yeah: xij and w **sum to 100**. **For EA and migshare, each zone is different and can sum to 100.**

Just looking at the patterns it's pretty dubious to say anything about employment but hey.

So I think for now, just concentrate on getting tables of everything out then think what to do next.

OK, I am googling shit about vars summing to one and getting nowhere again. Let's just get the tables out. So that requires running all the 3 census and having the numbers ready.

### 3 census numbers

Ah, I think the 3 census compile sheet function doesn't have the extra year adding code. Should be easy in this case, right? Maybe?

Wait no, cobtoregressionreadydata should be able to do it?

Oh right: **cobtoregressionreadydata is a subfunction of compileSheet...**

Actually, I think maybe it should already have all of the correct data...

Oh and I have some of the variable names wrong above. Giving them some sensible names might be an idea, huh? Oh actually no I've got them correct.

So what extra does the 3 census need?

* W and migshare for 2001.
* Density and crowding for 2001.
* Employment for 2001.
* House prices for 2001.

Which surely I had somewhere...? What's in the original eViewsFile? Amazingly, no, I've only got the 1991 numbers. Great.

So employment can just be shared between the two of them. And population count, in theory, though it should prob come from the original numbers.

What else was there? Back to eViewsReady... this is a peculiar kind of torture.

This is a fucking mess.

..

Plan: come back to 3 census 2001. Use what I've got, output everything. I can then come back and refactor this code properly as it's an absolute car crash. OK then.

Decision: I'm just going to leave rest of world out of the 5 census. For the rest, rich/poor split seems fine.

..

Getting there. A couple more things:

* Set estimates to zero if ... hmm, shall we say p < 0.05? Yah.
* Spread. Which I think should achieve what we want with some added formatting. (Needs only estimate being kept.)

OK, now to fuck about with formatting in Excel. Then maybe making graphs or maybe not.

### Monday bimbling about

Gwilym's got some tables but I'm not sure he'll get much of use from them. My thought for the day:

* Shares summing to 100: it just doesn't work if you're comparing different places. The numbers will be different depending on the number of zones, ceteris paribus.
* So comparing different cities is inherently problematic.
* I'm not sure this is solved by just sticking to 100% over the whole country, though at least that is consistent.

So. I should/could check those underlying numbers. Maybe I could do that right now, I have all the data.

..

OK actually: doing viz first. Can then actually show why it's weird, if it is. Need to nab code from course then adapt.

..

First pass OK. More I can do to make this clear. So staring at it:

* Rich/poor can probably be dodged.

Done that. What next? I think facetting based on type and making the order sensible, but let's think about that a bit...

* 5 vs 3 census cities. Best draw this.

I have a picture. I think I need to put the year in its own column... or rather, 3 censes, 5 census 71, 5 census 91. And completely remove numbers from the coeff codes.

..

Getting there. Now:

* Factor up things to get the right order.
* Add title and loop over coefficients.

Does it work for selecting ones that only appear in certain sources? Dunno, let's find out.