Estimating and profiling missing housing and population data for Ireland

Seán Healy, Soren Dreano
{healys47,soren.dreano2}@mail.dcu.ie [*]

November 31, 2021

Abstract

This work models missing historical and projection data for Irish housing and population on the county level, using available property price and population data. Analysis was carried out in order to determine statistical correlates. Then linear regression was used for the purpose of determining missing historical property prices, missing population projections, all on the county level. Finally, for visualisation purposes, county rent profiles were produced using a clustering technique.

Keywords: accommodation, housing, population, rent

1 Introduction

The CSO provide a very extensive collection of data sets at `data.cso.ie`. These data cover rent and property price, population records, estimates and projections.

Despite the broad array of datasets, it remains difficult to statistically compare variables, since the datasets tend to vary in geographic granularity, interval granularity, and overall timespan.

In ascending order of size, the geographic granularities (or zone type) are **eircode prefix**, **sub-county** (e.g. Dublin 2 or Fingal), **city**, **county**, **province**, **region** and **state**. Online resources provide a mapping of these granularities to each other, e.g. eircodes are mapped within counties via crowdsourcing [1], as listed in Appendix C. Counties are then mapped to geographic regions by the CSO itself [2]. The remaining **state** level is an aggregate of variables across the entire Republic of Ireland.

Alongside this array of geographic granularities, there are a number of interval granularities used by CSO datasets. Some sets use **monthly** datapoints, others use **quarterly**, **biannual**, **annual** or **sparsely annual**. Census data is generally sparsely annual, for example, since there are only censuses every few years.

Granularity is not the only incompatibility across data sources from the CSO. The overall timespan for data sets vary greatly. Population records are split across multiple data sets, some dating as far back as 1841. By contrast, county population for 2006 is only released under a standalone dataset. Some rent data is released between the years 2008 and 2021, whereas county-level property price records only begin in 2010.

After aggregation is applied to datasets, rendering them compatible, linear regression can be used to estimate missing data when a strongly correlated variable is present. Then a third issue arises: information overload. Much of the counties near Donegal, for example, follow similar trends for rent. This makes data difficult to visualise and reason about. For this, "rent profiles" can be created via clustering. Finally, rent, population and property prices can be compared across these few profiles rather than across the overly granular county-levels, or the overly geographic CSO regional level.

The issues that this work attempts to address can be summarised under three headings, which are explained across sections 2.1 to 2.3.

2 Methodology

2.1 Aggregation and interpolation

Other than the regional and eircode mappings, which were compiled by hand from online sources [1, 2], all data were downloaded via the `data.cso.ie` website, and preprocessed to remove out-of-scope data. A listing of the CSO data source IDs, and what purpose they were used for in this work, may be found in Appendix B.

As suggested in Section 1, most data sources differed in granularity and timespan.

Source	*Interval*	*Zone type*	*Begins*	*Ends*
E2004	Sparse	County, city	2011	2014
C0103	Sparse	Various	2006	2006
B0102	Sparse	Various	1841	2002
PEC08	Sparse	Various	2011	2046
RIH02	Biannual	Various	2006	2021
HPM04	Monthly	Eircode prefix	2010	2021
HPM09	Monthly	Various	2005	2021
CIA02	Annual	Region, county	2000	2018
BHA12	Annual	Region, county	1975	2020
HSA09	Annual	State	1975	2016
BBM02	Monthly	State	1975	2008
PEA18	Annual	State	1987	2021

Table 1: Data sources considered or used in this work

When there were various zone types present in a data source (e.g. state, province and county), much of the data was fil-

[*] This work was completed as part of the CA-660 module.

tered so that only one zone type remained, ideally county data. To aggregate sum data, a simple sum of constituents was used. E.g. the total volume of house sales in a county is the same as the sum of volumes for that counties' constituent eircode prefixes.

Mean data was aggregated using a weighted mean. E.g. the mean property price in Louth is the volume-weighted mean of the house prices in eircode prefixes A91 and A92 (Laois' only eircode prefixes). This ensures an eircode with very little sale volume does not have a disproportionate effect on the aggregated mean for the larger surrounding region.

For pre-processing data into less granular intervals, a similar aggregating approach was used. Turning monthly sum data into annual sum data involves summing all the months' data points. In contrast, turning monthly mean data into annual mean data involves taking a weighted mean of the monthly data. E.g. in (1) the mean annual property $\mu_{\text{price}}(y)$ can be determined by iterating through each month m in year y, summing the product of the month's mean price $\mu_{\text{price}}(m)$ and the month's total volume $v(m)$. That entire sum is divided by the year's total volume $v(y)$, producing the year's mean property price.

$$\mu_{\text{price}}(y) = \frac{\sum^{m \in M(y)} \mu_{\text{price}}(m) \times v(m)}{v(y)} \tag{1}$$

Interpolation was used to convert from sparse to dense data. Census data is only available every few years, so in this work, linear interpolation was used to estimate population on state, county and region levels between any two given years, e.g for the 4 missing years between the 2011 and 2016 censuses. Unlike aggregation, which produces an exact, known value, interpolation introduces uncertainty, because the interpolated data points are only estimates. It was noted early on in this work that Irish population tends to move slowly and smoothly, so we chose linear interpolation to find missing points between known points, and accepted the inaccuracy of the estimates as negligible.

Interpolation isn't to be confused with the techniques that will be discussed in Section 2.2, where instead, two variables are considered, one known and another unknown. Interpolation uses one variable, and can only estimate points between two known points, i.e. nothing before the first or after the last point.

2.2 Estimating missing data

As mentioned in Section 1, linear regression was chosen as a candidate for estimating missing data points. For this to work, however, a strong correlation must first be found between two or more variables. Section 2.1 laid the foundation for comparison of different variables; instead of several incompatible tables, this section assumes pre-processed tables, where the interval granularity is annual, the zone type is mostly county, and missing years have been interpolated into usable data points.

County-level population projections were obtained by leveraging regression to estimate county-regional population proportions as a function of time, then combining the resulting estimates with CSO-published regional population projections.

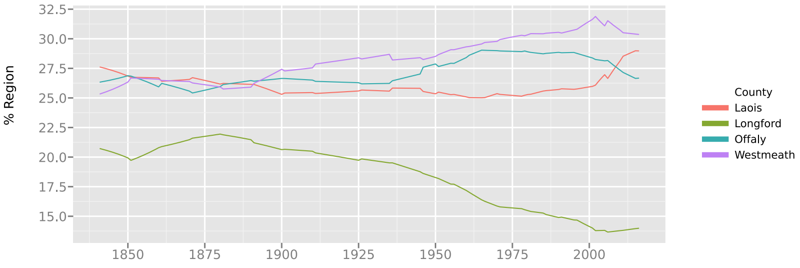

Figure 1: Distribution of population in the Midlands region

CSO presently only offers population projections for the broad regions outlined in Appendix A. However, it's possible to estimate the population projections for the counties that comprise these regions by splitting each region's projection in an uneven manner that reflects the population proportions counties hold in each region. Dublin is its own standalone CSO region, so no action was required in that case. However, in the case of CSO's Midlands region, for example, the population projections for the counties within that region could be estimated by splitting the regional estimate.

A naive approach to regional-to-county population projection conversion would assume that the proportion of a region's population that a particular county covers remains constant; if a region with only two counties, A and B, has a population ratio of $1 : 2$ between those two counties in 2021, the naive assumption would be that those counties would display the same ratio in 2036. However, it is observed that counties can become more or less prominent portions of their regions as time goes on. This is evident in Figure 1. Laois used to be the most populous county in the Midlands, but now Westmeath is, and Longford has become significantly less proportional in that region over the past century.

Using data from Figure 1, lines are fitted over the county plots in order to predict what regional proportion a county will comprise further into the future (e.g. in 2036). These lines are also normalised at each time interval (each year), so that the county proportions for each region sum to 1.0 at each interval.

When these predicted proportions are multiplied by the regional predictions, estimates for county populations are obtained. Figure 2 illustrates the efficacy of this model. The figure uses the model to produce estimations of pre-2017 population, which can then be compared against the known or interpolated populations for those counties at given years.

$$p(y, c) = p(y, r(c)) \times \frac{p(y, c)}{p(y, r(c))} \tag{2}$$

$$\hat{p}(y, c) = p(y, r(c)) \times \hat{q}(y, c) \tag{3}$$

The formula representing this regression technique is derived from (2). Here, $p(y, c)$ means *the population of* county c in year y. $r(c)$ returns the broader region of county c; e.g. $r(\text{'Donegal'}) = \text{'Border'}$. The observation (2) has one particular part, $\frac{p(y,c)}{p(y,r(c))}$, for which an estimate $\hat{q}(y, c)$ has previously been suggested. $\hat{q}(y, c)$ is defined as the predicted proportion of the outer region a county c fills during year y. E.g.

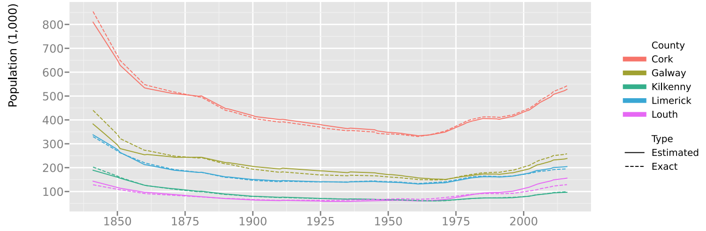

Figure 2: Regression county population estimation

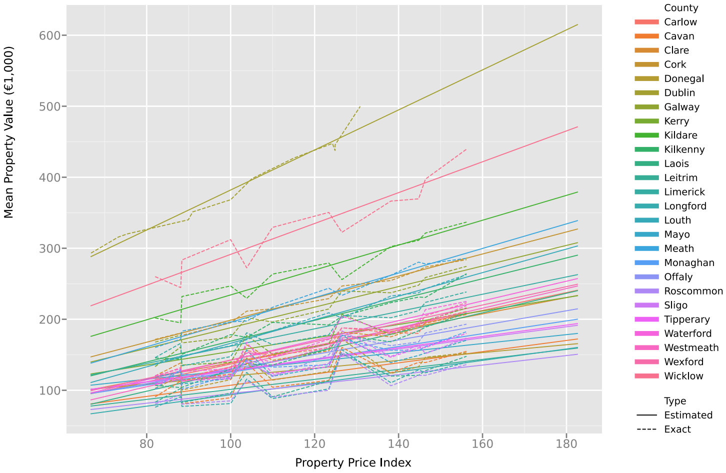

Figure 3: Estimating mean property price using linear regression over CSO price indices

$\hat{q}(2017, \text{`Westmeath'}) \approx 30\%$, and $\hat{q}(2036, \text{`Westmeath'}) \approx 33\%$. This can be described in terms of *Westmeath becoming increasingly populous within the context of its Midlands region*. Finally, (2) leads to a model (3) for estimating county population when only regional population predictions and historical county population records exist.

Regression was also used to estimate missing historical data for house prices. CSO publishes property price data in two ways: firstly a historical price 'index', and secondly, exact sale volumes and mean prices. The problem is the former data source is broader in time range the latter. Mean price data is missing before 2010, whereas the index data goes as far back as 2005. From a research perspective, it would be useful to obtain estimated house price data that spans before the 2008 recession.

After verifying that the index does indeed correlate with the price data during overlapping years, linear regression was used to extend estimated mean property price data back to 2006. One linear regression pass per county was used, as illustrated in Figure (3).

2.3 Profiling

After using linear regression to estimate missing data (both in the past and future), a problem of information overload became more apparent. Some counties correlated so closely in terms of rent prices that any study of rent in Ireland would benefit from forming *profiles* or *county groups*. A form of hierarchical clustering was used to group counties into 5 profiles based on historical rent prices. The heuristic used was *lowest difference between rent as a function of time*, $\Delta(g_1, g_2)$. This approach is outlined in Algorithm 1, which takes a set of counties C as input, along with a desired number of rent profiles, k. The counties are first converted to singleton profiles $\{c\} \in K$. At initialisation $|K| = |C|$, but each round, a profile g_1 is merged with another, g_2, according to minimum difference ($argmin_{\{g_1,g_2\} \subset K} \Delta(g_1, f_2)$), reducing the cardinality of $|K|$ by one. This repeats until we are left with the desired k rent profiles.

Algorithm 1 Clustering counties by rent similarity

```
Input: $C$, $k$
$K \leftarrow \{\{c\} \mid c \in C\}$
while $|K| > k$ do
    $\{g_1, g_2\} \leftarrow argmin_{\{g_1,g_2\} \subset K} \Delta(g_1, g_2)$
    $K \leftarrow (K - \{g_1, g_2\}) \cup \{g_1 \cup g_2\}$
```

3 Results

As illustrated in Figure 4, the resulting rent profiles from Algorithm 1 depart from the standard geographic regions used by the CSO. These profiles (outlined in Table 2). may provide a better lens for comparison of rent progression and property price progression going forward.

Group	*Rent*	*Counties*
1	Low	Cavan, Roscommon, Donegal, Longford, Leitrim, Mayo, Tipperary
2	Medium-low	Carlow, Kilkenny, Clare, Kerry, Wexford, Waterford, Offaly, Sligo, Laois, Westmeath
3	Medium	Cork, Galway, Meath, Limerick, Louth
4	Medium-high	Kildare, Wicklow
5	High	Dublin

Table 2: Rent groups

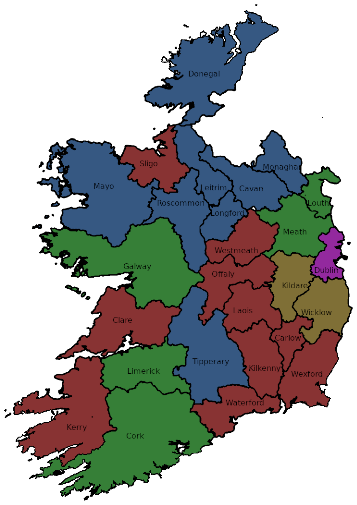

Figure 4: Rent profiles

These profiles were used to produce the historical property price plot in Figure 5, which also includes estimated prop-

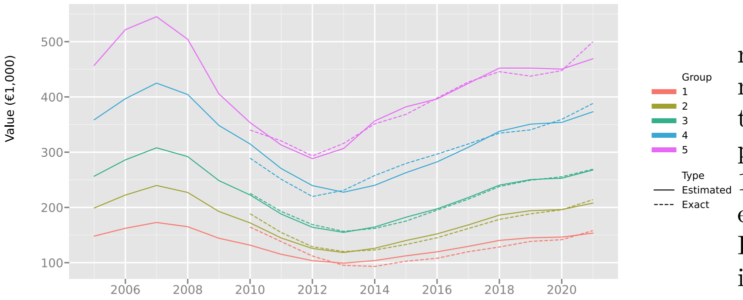

Figure 5: A fuller picture of property prices in Ireland

erty prices resulting from the methods outlined in Section 2.2. Similarly, Figure 6 demonstrates how total populations are expected to rise in the various rent profile groups over the coming years. The figure also includes historical population for comparison. This figure was produced using the regression techniques from Section 2.2, since CSO only provide regional population projections.

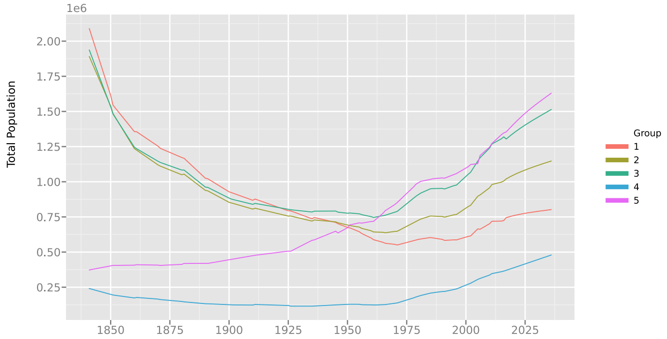

Figure 6: Rent profile population history and projections

4 Discussion and Conclusion

The mean price to buy a property decreased from 2010 to 2012 after the Irish Property Bubble collapsed. The price later rose in all county groups, as shown in Figure 5.

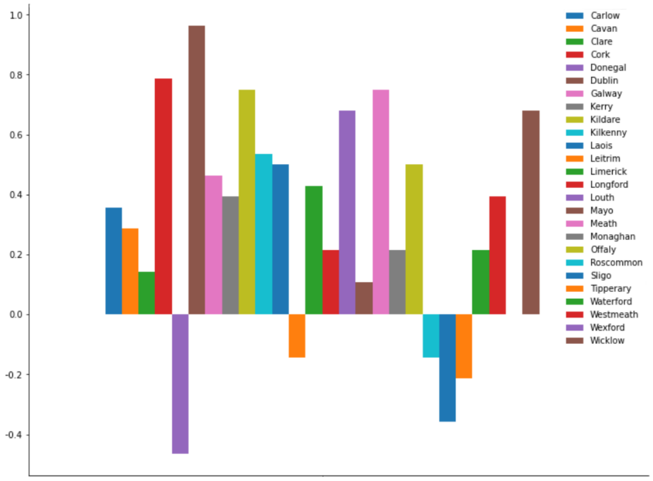

Figure 7: Correlation of population and prices in counties

Increase in inhabitants might drive prices in the housing market up, as there will be more demand for limited accommodation. In order to test this theory, we looked at correlations between the total population of a county and the mean property price, from 2010 to 2016, for each county. Group 1 (low rent) actually has negative correlation between property price and population, as illustrated in the columns for Donegal, Leitrim, Roscommon, Sligo and Tipperary, meaning property prices decreased while the population continued to increase in all of the aforementioned counties.

In the later part of the Celtic Tiger, the Irish Property Bubble started to appear, which manifested itself as a constant and continuous rise in prices for almost a decade. In 2008, the Irish government acknowledged the country's descent into recession [3], the collapse of the previously mentioned bubble being one of the major contributing factor to the global 2008 financial crisis. Unemployment rate rose from 6.5% in July 2008 to 14.8% 4 years later [4] and house prices fell from 35% between the end of 2007 and the end of 2010 [5].

Current Tánaiste Leo Varadkar recently said "One of our biggest deficiencies, in housing supply in Ireland, is we're a country of three-bed homes by-and-large and we don't have enough one-bed homes" [6]. We could not find any publicly available data on the matter. Nonetheless, an analysis on the housing supply in Ireland either by size or the number of bedrooms, especially in dense urban areas such as Dublin, Cork or Galway might confirm this deficiency. Further research should focus on the types of housing that sell the most and that have been constructed recently.

References

[1] Unknown, "Croudsourced eircode map," 2019, http://eircode.codes/.

[2] CSO, "Regional population projections 2017 - 2036," 2016, https://www.cso.ie/en/releasesandpublications/ep/p-rpp/regionalpopulationprojections2017-2036/appendix2conceptsanddefinitions/.

[3] J. Kollewe, "Ireland falls into recession," *The Guardian*, 9 2008, https://www.theguardian.com/business/2008/sep/25/recession.ireland.

[4] CSO, "Seasonally adjusted standardised unemployment rates (sur)," 2014, https://www.cso.ie/en/statistics/labourmarket/principalstatistics/seasonallyadjustedstandardisedunemploymentratessur/.

[5] H. Environment and L. Government, "Quarterly house prices bulletin," *Environment, Heritage and Local Government*, 2010, http://www.environ.ie/en/Publications/StatisticsandRegularPublications/HousingStatistics/FileDownLoad,25191,en.pdf.

[6] D. McGrath, "Young people want one-bedroom apartments, not three-bed homes," *The Irish Times*, 11 2021, https://www.independent.ie/breaking-news/irish-news/young-people-want-one-bedroom-apartments-not-three-bed-homes-varadkar-41089264.html.

A CSO Regions

Region	*Counties*
Border	Donegal, Sligo, Leitrim, Cavan, Monaghan
Dublin	Dublin
Mid-East	Wicklow, Kildare, Meath, Louth
Midlands	Longford, Westmeath, Offaly, Laois
Mid-West	Clare, Tipperary, Limerick
South-East	Waterford, Kilkenny, Carlow, Wexford
South-West	Cork, Kerry
West	Galway, Mayo, Roscommon

B CSO Data Sources

Region	*Counties*
Population	E2004, C0103, B0102
Population projection	PEC08
Rent	RIH02
Property price	HPM04
Estimated property price	HPM09
Income data	CIA02
Planning permission data	BHA12
Construction cost data	HSA09
Construction employment data	BBM02
Migration data	PEA18

C Eircode Prefixes

County	*Eircode Prefix*
Carlow	R21, R93
Cavan	H12, H14, H16
Clare	V14, V15, V95
Cork	P12, P14, P17, P24, P25, P31, P32, P36, P43, P47, P51, P56, P61, P67, P72, P75, P81, P85, T12, T23, T34, T45, T56
Donegal	F92, F93, F94
Dublin	A41, A42, A45, A94, A96, D01, D02, D03, D04, D05, D06, D6W D07, D08, D09, D10, D11, D12, D13, D14, D15, D16, D17, D18, D20, D22, D24, K32, K34, K36, K45, K56, K67, K78
Galway	H53, H54, H62, H65, H71, H91
Kerry	V23, V31, V92, V93
Kildare	R14, R51, R56, W12, W23, W34, W91
Kilkenny	R95
Laois	R32
Leitrim	N41
Limerick	V35, V42, V94
Longford	N39
Louth	A91, A92
Mayo	F12, F23, F26, F28, F31, F35
Meath	A82, A83, A84, A85, A86, C15
Monaghan	A75, A81, H18, H23
Offaly	R35, R42, R45
Roscommon	F42, F45, F52
Sligo	F56, F91
Tipperary	E21, E25, E32, E34, E41, E45, E53, E91
Waterford	X35, X42, X91
Westmeath	N37, N91
Wexford	Y21, Y25, Y34, Y35
Wicklow	A63, A67, A98, Y14

D Property price rise since 2008

County	Increase in property price (%)
Dublin	36.53
Kildare	36.49
Wicklow	34.43
Cork	20.37
Galway	11.44

E Code repository

https://github.com/Gailenstorm/CA-660