Modelled Territorial Authority Gross Domestic Product estimates for New Zealand

Peter Ellis

Ministry of Business, Innovation and Employment; views expressed are the author's.

June 12, 2016

Abstract

Since late 2015 the Ministry of Business, Innovation and Employment has been publishing estimates of Modelled Territorial Authority Gross Domestic Product. The latest available estimates have detailed industry breakdown to 2013 and total GDP estimates to 2015. Per capita and real estimates are provided as well as nominal. The estimation combines official statistics from the business demography statistics, Linked Employer-Employee Database, 2013 Census, Regional GDP and National GDP. Statistical methods used include iterative proportional fitting and time series forecasting techniques. The data are available via two distinct interactive web tools and as a bulk download for re-use. All the source code and data used in creating the estimates have been published. An example analysis of the impact of economic conditions in 2005 on subsequent real GDP growth, taking into account spatial distribution of error terms, is provided in this paper.

List of Figures

1 Motivation

Since 2014, the "Regional Gross Domestic Product (RGDP)" official statistics series,[29] funded by Vote Economic Development but developed and published by Statistics New Zealand, estimates GDP for Regional Council areas according to a high level industry classification (ie 15 regions by 17 industries). However, most regional economies and polities in New Zealand comprise both urban and rural districts. There remains a demand for finer scale information (both geographically and for industry classifications) for better understanding the situation and trends *within* Regional Council areas.

In recent decades a range of Territorial Authority estimates of economic activity have been in use, but these have for the most part (perhaps entirely) been commercial products from specialist economic consultancies. The details of the methods used to prepare estimates have not been published, nor have the estimates themselves. Analysts and officials wanting to build on previous economic analysis have found themselves in the position of paying multiple times to access a single set of data. Critique and improvement of methodology has not been possible because of the commercial-in-confidence nature of the process.

Further, to my understanding (which cannot be confirmed for the reasons set out in the previous paragraph), these commercially available estimates did not reconcile to the officially published Regional or National GDP figures from Statistics New Zealand.

In 2014, the Ministry of Business, Innovation and Employment (MBIE) initiated a "Modelled Territorial Authority Gross Domestic Product" project to address these issues.[16] It had the following objectives:

- Publish estimates of GDP at the Territorial Authority level consistent with the Regional and National GDP Tier 1 statistics
- Make the full set of data available to researchers and policy-makers for re-use; freeing up resource for analysis rather than purchase and re-purchase of data
- Make the source code and detailed methodology available for critique and improvement
- Release the data for non-specialist users in interactive web tools to facilitate general usage and informing of policy debate

2 Method

2.1 Overview

Estimation of MTAGDP is a two stage process. Statistics New Zealand's RGDP estimates of Gross Domestic Product for 15 Regions (and regional groupings) for years ending in March are a key source, and the two stages of the method reflect the levels at which these statistics are available. RGDP is published at a total GDP level to the year of publication minus one; and at a high level industry breakdown to current year minus three.

2.2 Scaling up to industry breakdowns

The first stage of estimation works to the RGDP industry totals for current year minus three, which is 2013 at the time of writing. The data sources used at this stage are:

Business Demography Statistics [28] – two-way table of employee numbers, Territorial Authority by fine level (ie 6 digit ANZSIC)

Linked Employer–Employee Data Table 4 [27] – Earnings by Quarter and fine level (ie 6 digit ANZSIC06) industry classification

Linked Employer–Employee Data Table 37 – Earnings by Quarter and Territorial Authority

Linked Employer–Employee Data Table 18 – Earnings by Quarter and Region, and medium level industry (ie 3 digit ANZSIC)

Custom data table from Statistics New Zealand for Regional GDP – providing similar geographic resolution to published RGDP (15 regions) and finer industry classifications (30 industries).

National Gross Domestic Product Production measure, nominal – provides more detailed industry breakdown than GDP but no regional information.

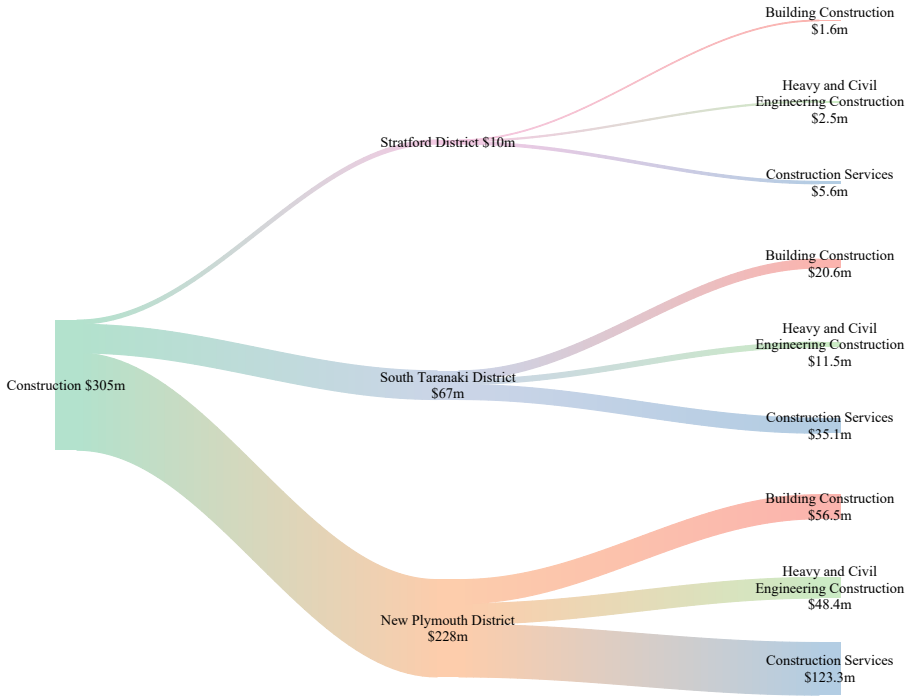
Regional Gross Domestic Product – provides regional GDP by industry up to 2013 and regional totals to 2015; national totals match (within rounding error) those in NGDP.

The method can be envisaged as taking the Business Demography statistics - which are at the necessary level of granularity (in fact, somewhat finer) but represent the wrong value variable - and weighting them up to regional and national GDP which have the correct value variable but not enough granularity. The quarterly earnings information from the LEED acts as a compromise intermediate step between the two.

Alternatively, the method could be seen as taking the RGDP totals and allocating them out to more detailed breakdowns of district/city and industry based on the earnings and employee numbers; making the minimal assumptions needed on consistency of the ratio of value add to earnings in given combinations of region and industry, and from earnings to employee numbers in given combinations of territorial authority and industry.

Technically, the employee numbers from the Business Demography Statistics are weighted by iterative proportional fitting so their marginal

Figure 1: Construction GDP in Taranaki and its Territorial Authorities in 2013



totals match the various marginal totals that can be derived from the three LEED tables. This provides estimates of earnings at the desired level of granularity. Then those estimates of earnings are weighted up to match the marginal totals derived from RGDP and NGDP; providing estimates of GDP or value added at the correct degree of granularity.

The resulting top-down aspect of the process can be visualised in Figure 1, a Sankey chart[21] which illustrates how the published RGDP figure for construction in Taranaki ($305 million in 2013) is allocated out to the Territorial Authorities that overlap that region, and also to the more detailed NGDP level of industry. This distribution is done on the basis of proportions of earnings from the LEED where they can be used to 'spread out' RGDP to a finer level; with those LEED earnings themselves spread out proportionate to number of employees when necessary.

The bulk of the work in the project was in managing concordances and classifications across the various data sources. For example, Territorial Authorities are not strictly hierarchically organised under Regional Councils. Rather, the two classifications have a many-to-many relationship that needed to be carefully handled. All the datasets had slightly different geographic and industry classifications.

Implementation of the project was made possible by the flexibility of

the R statistical computing environment[17] with extensive use made in particular of Wickham and Francois' `dplyr` package.[23] The iterative proportional fitting was done by re-purposing the `rake` function from Lumley's `survey` R package.[14]

2.3 Commuter correction

One limitation of using different employer–employee sources for deriving estimates of GDP at the TA level is that the business demography and GDP figures are based on place of production, whereas the earnings tables available from LEED are based on the employee's home address. For Territorial Authorities where there are a considerable number of work commuters across districts (eg Wellington City receives a large number of commuters from Lower Hutt City, Upper Hutt City, Porirua, Kapiti Coast, et cetera), this effectively means that the production-related earnings are (undesireably) transferred across Territorial Authority boundaries.

In order to correct for the transfer of earnings across Territorial Authority boundaries, data from workplace and home addresses from the 2013 Census were used to calculate the relative proportion of earnings based on the reported commuter numbers. Figure 2, drawn with the help of Csardi and Nepusz's `igraph` software,[5] illustrates these data.[6] This approach is not fully satisfactory (for example, it cannot address the obvious fact that some industries and occupation classes will have a higher proportion of commuters than others), and improving it is noted as an area of future work.

2.4 Inflation-adjusted and per capita measures

The published data include per capita estimates. Population totals by Territorial Authority came from Statistics New Zealand.[26]

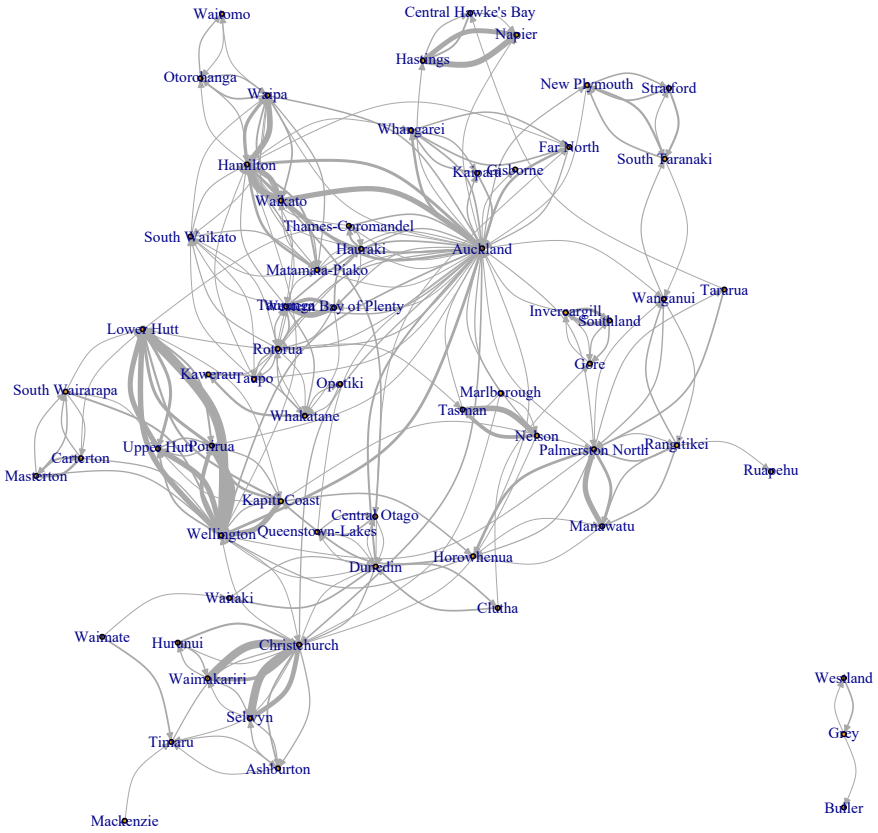
The published data also include inflation adjustments, in an attempt to create a measure of the volume of production. Due to limitations in regional pricing data the adjustments have been made identical across New Zealand. The deflators used were derived by comparing Statistics New Zealand nominal[31] and chain volume[30] series. This provided industry-level deflators for 31 industry categories.

2.5 Forecasting totals for two most recent years

The bottom up method described in the previous section requires industry level GDP estimates to weight up to. For the two latest years of RGDP only total GDP by Region is available. To provide equivalent estimates at Territorial Authority level we use time series method to 'forecast' total GDP by Territorial Authority for the two final years, and then a final use of iterative proportional fitting to weight up those forecasts to match the published regional totals. In effect, the forecasts are of trends in Territorial Authorities' share of a region's total GDP, and the forecast share is applied to the published total.

To make the short term forecasts as smooth and plausible as possible, the annual series of total GDP by Territorial Authority derived in the first

Figure 2: Commuting patterns in 2013 (only paths with > 100 people shown)



stage of the modelling process are decomposed into smoothed monthly figures. "Temporal disaggregation" is a technique to disaggregate low frequency time series to a higher frequency. We used the Denton-Cholette method as implemented in Sax and Steiner's `tempdisagg` R package.[18]

The forecasts themselves are done using Hyndman et al's `hts` R package,[9] which greatly facilitates time series forecasts of hierarchical and grouped time series.

2.6 Assumptions and limitations

The following can be regarded as the key assumptions made in the allocation of published GDP to finer levels of spatial and industry granularity.

2.6.1 Earnings a good indicator of total value added

In a very general sense, we rely on wage and salary earnings from the quarterly LEED data being a good indicator of total value added which makes up GDP. Most obviously, profits and self-employed earnings are not included in the quarterly LEED data. We rely on the ratio of wage and salary earnings being consistent within a particular industry - region combination.

2.6.2 No interaction between TA, and earnings to GDP ratio in a given industry

More specifically, the Regional GDP Tier 1 statistic gives GDP for a given industry in a particular Regional Council. To allocate that GDP to the Territorial Authorities that share space with that region, we need to assume that the ratio of earnings to GDP in that industry is the same in each of its Territorial Authorities. This is clearly implausible because there is almost certainly a spatial element to profit to wage ratios, but we have to hope that it provides a reasonable approximation.

2.6.3 No interaction between Region and TA, and the national inter-detailed-industry relative earnings to GDP ratios

National GDP confronted with the LEED provides reliable earning to GDP ratios for the 56 industries available in the National GDP series. Regional GDP confronted with LEED provides such ratios for 17 higher-level industries at the Regional Council level. To allocate GDP to the more detailed industries in Regions and Territorial Authority we need to assume not that the ratios in each region match the national ratios, but that the ratios for the detailed industry that make up a single higher level industry are the same *in relative terms* around the country.

2.6.4 Employee numbers a good indicator of earnings

For the earnings to GDP ratios mentioned above to mean anything, we need earnings estimates for the correct combination of variables. When earnings data are not available at the required granularity (combination of Territorial Authority and 56 industries), but are available one level up (eg by Region and higher level industry) we rely on employee numbers from the Business Demography Statistics to allocate the earnings to the more detailed level while constraining all available LEED marginal totals to be met. In effect, we rely on the ratio of employees to earnings being similar across a range of classification boundaries: Regional Council to Territorial Authority, and high level industry to more detailed industry. This is a significant assumption across the detailed industry classifications and a potential area of improvement if a better data source could be found.

2.6.5 Commuting patterns are not industry specific

Because the LEED data on earnings are only publicly available based on place of residence (not place of work), we use commuting patterns reported in the 2013 census to shift a proportion of all earnings by Territorial Authority into the destination districts and cities to which commuting is reported. We have no data readily available to allow this to be done on an industry basis, so the implausible assumption that commuting patterns are the same in each industry is required. One would expect that local service industries like hairdressing and small scale retail (ie the corner dairy) would experience less commuting than financial and professional services and public administration. But data to improve on this assumption still needs to be integrated into the project.

2.6.6 Commuting patterns haven't changed over time

For convenience, we used only the 2013 Census for the "commuting correction" mentioned above, and applied its correction factors to all years from 2000 to 2013 on an implausible assumption of stability in commuter patterns. Clearly this could be improved on by using similar results from the 2000 and 2006 censuses for earlier years.

2.6.7 Price movements by industry are not region-specific

To produce estimates of GDP in real terms it was necessary to use industry-level price series derived from National GDP, implicitly assuming that price movements are not region-specific. This is clearly not going to be correct, but better than no inflation adjustment at all.

3 Results

3.1 Overall results

Figure 3 illustrates high level results of the whole exercise, showing estimated real economic growth per person over ten years. Some of the highlights of that figure are doubtless surprising, so it is complemented with Figure 4 which contrasts GDP growth with growth per person. The difference of course is purely population movements, but the figure nicely highlights a few points of interest. For example, Selwyn and Queenstown are estimated to have experienced respectable economic growth in absolute terms (4.4% and 2.3% respectively) but in both cases population growth has kept pace. We might speculate that the causality is not necessarily in the same direction in each case.

The apparent success of Opotiki (to pick just one example) might come as a surprise to some readers, so Figure 5 provides an industry breakdown at the same level as provided in the Regional GDP. Driving the estimate of the growing GDP in Opotiki is the growth in earnings reported in the LEED. At an industry level, it is services industries that have apparently seen strong growth.

Data of this sort aggregated by statistical or political boundaries lends itself to representation in a choropleth map, so Figures 6 and 7 illustrate this. Of limited value analytically, such maps are useful for raising awareness of results and of the mere existence of data.

Visually very similar despite covering different time periods and different slices of data (one showing total GDP and one just construction), Figures 6 and 7 tell a joint story of economic growth stronger in the South Island than in the north over the space of a decade or so.

Figure 3: Average real growth in per person GDP 2005 to 2015

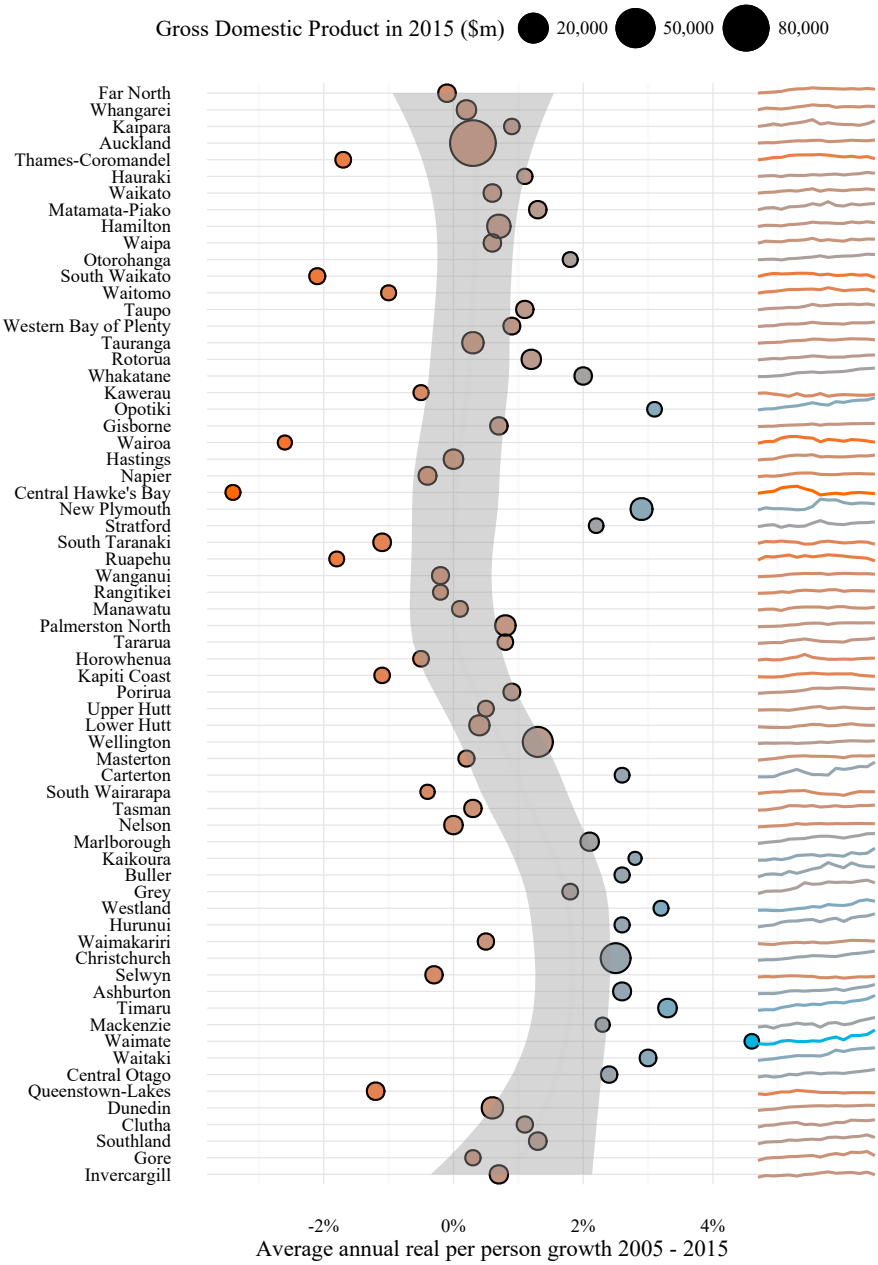


Figure 4: Average growth rates 2005 to 2015: per person and absolute

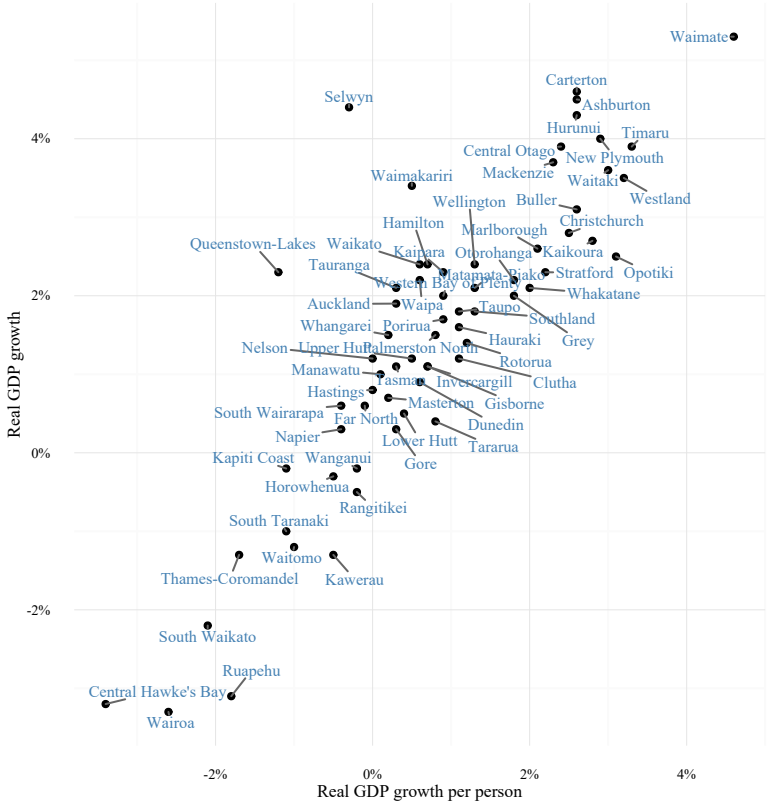


Figure 5: Real GDP growth in Opotiki by high level industry

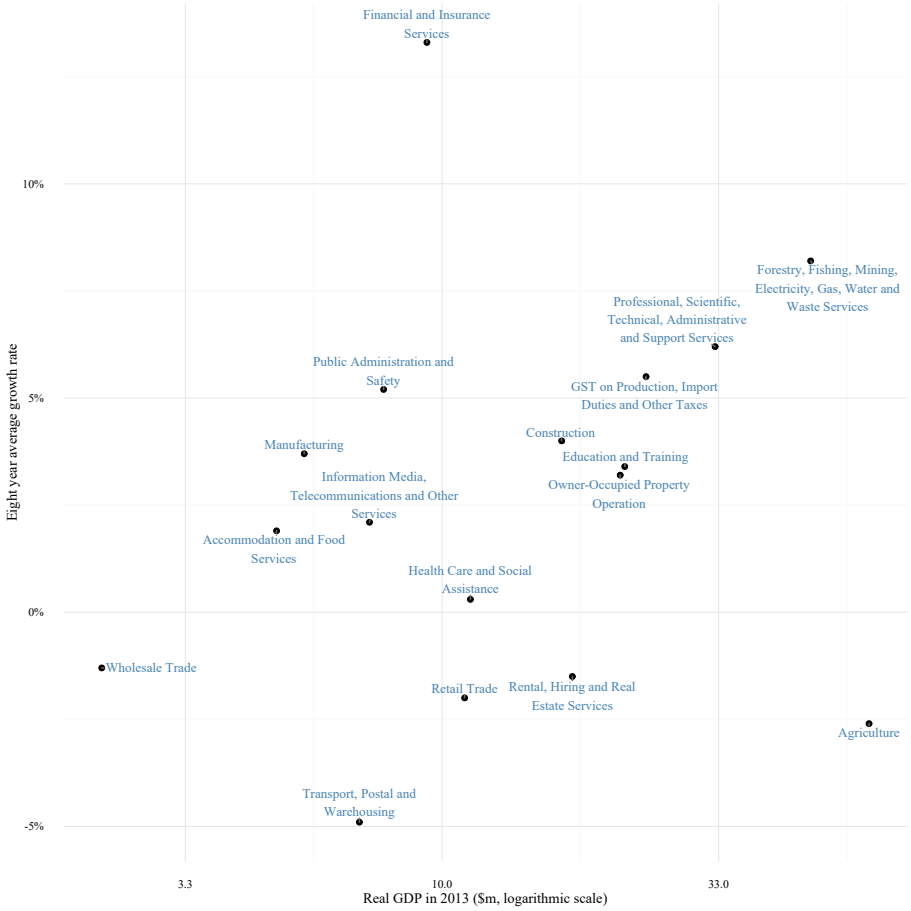


Figure 6: Average real growth in per person GDP 2010 to 2015

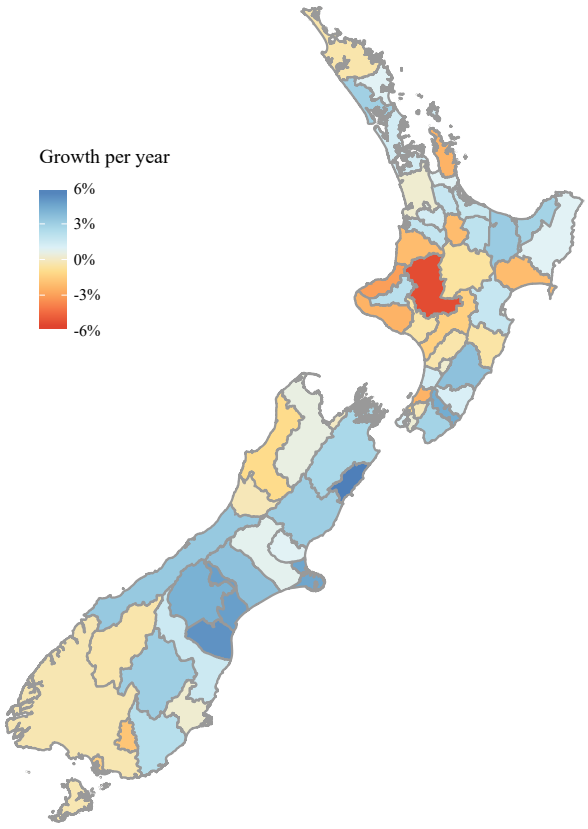
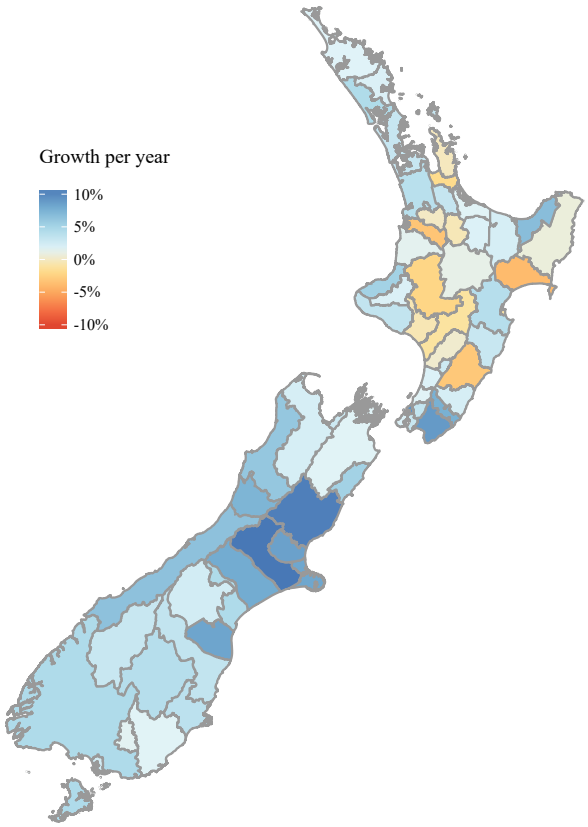


Figure 7: Average real growth in construction GDP 2003 to 2013



3.2 Cluster analysis

For any given time period, it is possible to construct a matrix of GDP values by industry and Territorial Authority and hence to apply multivariate statistical techniques to identify patterns. To illustrate this potential use of the data, I have constructed a 66×56 matrix where each row represents a Territorial Authority and each column a detailed (National GDP classification) industry. The values in cells are the estimated Gross Domestic Product in 2013. The *rows* have been scaled to have a mean of zero and variance of one, so each Territorial Authority is treated as of equal weight.

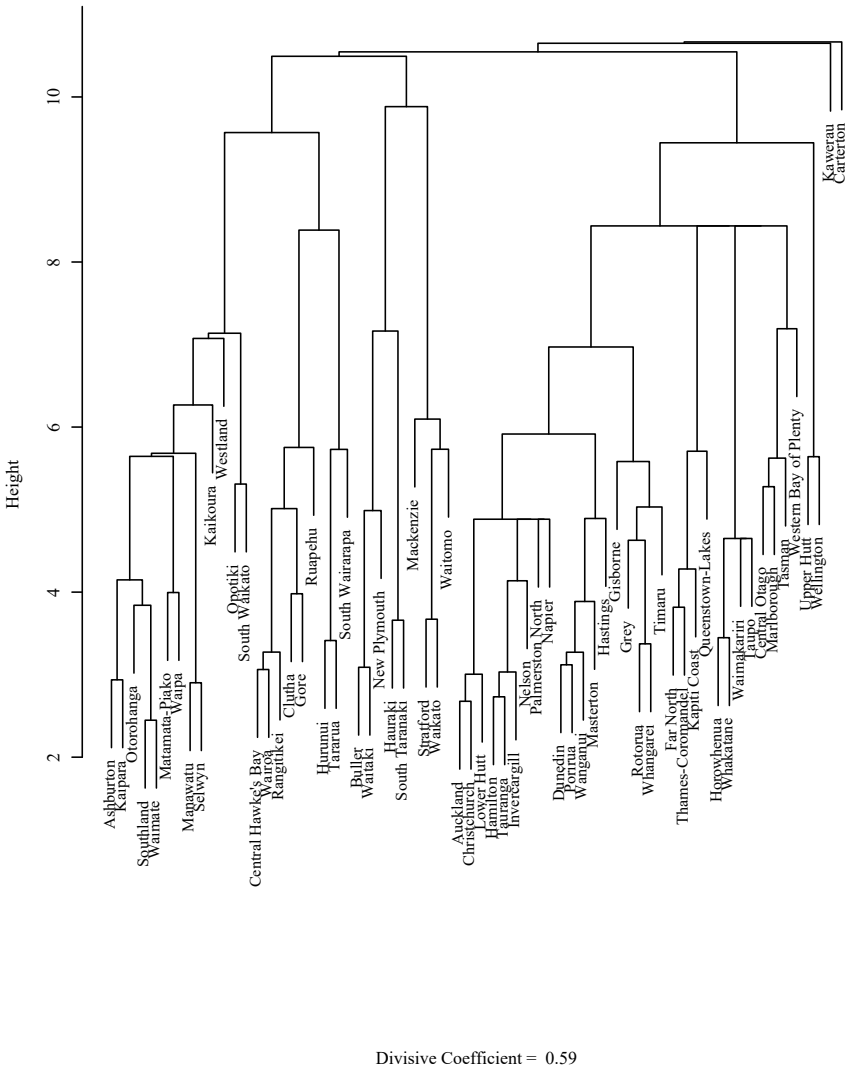
Figure 8 applies DIvisive ANAlysis Clustering ("DIANA")[13] to that matrix to identify Territorial Authorities that have similar industry profiles. DIANA is a divisive hierarchical clustering method that starts with one large cluster of all the observations, and divides it until each cluster contains only a single observation. To divide a selected cluster, the algorithm seeks its most disparate observation, which initiates a "splinter group" and takes away with it similar observations (in this case Territorial Authorities with similar industry profiles). The result can be visualised in terms of a tree as in Figure 8. DIANA is implemented in R in the `cluster` package.[15]

Figure 8 shows some interesting associations. The first splinter group constitues Kawerau and Carterton - two districts similar to each other in profile, despite lack of proximity. Both are rural districts with a noticeable localised manufacturing sector.

Other clusterings will be unsurprising; for example, at the bottom left of Figure 8 we see a group of districts - Ashburton, Kaipara, Otorohanga, Southland, Waimata, Matamata-Piako, Waipa, Manawatu and Selwyn - spread around the country but with strong dairy sectors. The easiest way to confirm what these districts have in common is via the "One area's top industries" tab of the MTAGDP interactive web tool, to be described later in this paper.

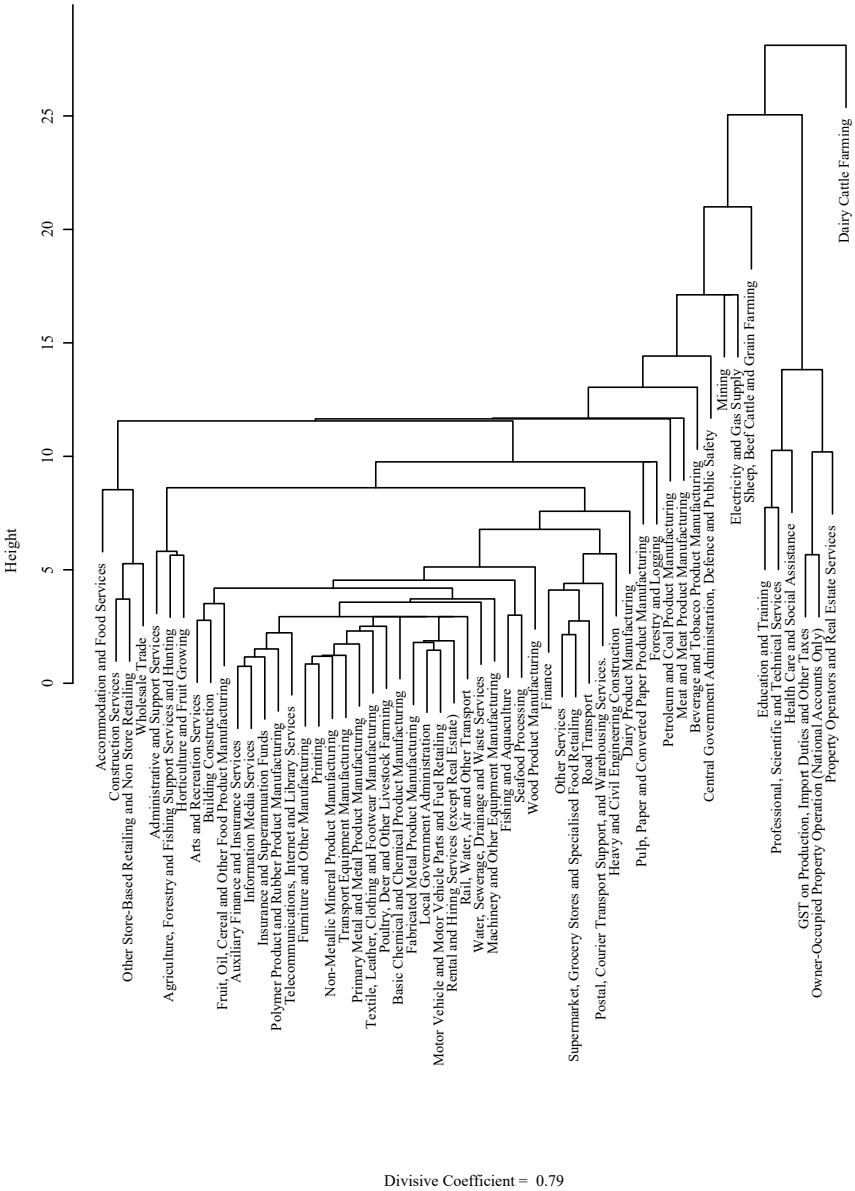
Figure 9 applies the same DIANA algorithm to the transposed version of the matrix described above. It provides insight into which industries tend to be co-located. Dairy Cattle Farming stands out on its own and provides distinction to any district that focuses on it, but other industries are more closely related. For example, "Mining" tends to be associated with "Electricity and Gas Supply". Perhaps more interestingly, towards the left of the diagram, "Construction Services" has an association with "Wholesale Trade" and "Other Store-Based Retailing and Non Store Retailing". This does not necessarily indicate a direct economic link between the industries, but illustrates a spatial association that is most likely to do with patterns of settlement and commercial services agglomeration. The MTAGDP data should provide a rich resource for economists studying these economic geography questions in New Zealand.

Figure 8: Similarities of Territorial Authorities based on industry



Divisive Coefficient = 0.59

Figure 9: Geographical associations of industries



Divisive Coefficient = 0.79

3.3 Example analysis - agricultural concentration and GDP growth

In this section I work through an end-to-end example analysis of the MTAGDP data in the hope that it will motivate others to attempt more ambitious analyses of substantive questions and integrate the data into ongoing research programs.

3.3.1 Questions and data

I chose to investigate any relation between two economic variables observeable in 2005 and growth in real GDP over the subsequent ten years. The response variable is summarised in Figure 10. The two explanatory variables of interest are:

- the proportion of districts' and cities' GDP that came from agriculture in 2005
- nominal GDP per capita in 2005

Between them, these variables will help us see (from one angle) if the industrial focus in 2005 contributed to subsequent growth; and if the state of prosperity in 2005 contributed to subsequent growth either positively ("the rich get richer") or negatively (perhaps some kind of regression to the mean). 2005 was chosen as a stable reference year prior to the instability of the 2007 - 2008 global financial crisis, and as facilitating a ten year longer view of New Zealand's economic story. The data used are the MTAGDP data by Territorial Authority and "RGDP_industry" ie the industry classification used in the Regional GDP official statistic.

I limited myself to two variables of interest due to the relatively small dataset. Following model-building strategies set out by Harrell,[10] with observations on 66 Territorial authorities, we have between 3 and 6 degrees of freedom to allocate to parameters before we are irretrievably overfitting. From previous experience with similar spatial data I suspect 3 or 4 effective degrees of freedom (at least) will be need to take into account spatial patterns in the randomness, leaving two fixed variables of interest my likely maximum.

These data can be envisaged as originally longitudinal or panel data. The basic relationship is shown in Figure 11. However, to simplify the modelling challenge for demonstration purposes, I reduce the longitudinal time series aspect of the data to a single ten year real GDP growth rate used as the response variable. I chose absolute growth for the response variable rather than growth per person because it is likely that population movements will follow economic success and failure - as we have already seen in the case of Queenstown. In contrast, GDP per person at the snapshot moment of 2005 makes sense as an explanatory variable, standing in as a proxy for prosperity at the beginning of the period.

The two key bivariate relationships of both agricultural focus and GDP per person in 2005 to subsequent economic growth is seen in Figure 12, where we reduce growth from 2005 to 2013 to a single figure. The robust lines of best fit shown are from a Huber M-estimator.[20]

Figure 13 is the spatial equivalent of Figure 12 in that it shows the relationship between latitude and longitude and 10 year economic growth, disregarding the other two variables in the model. The colouring and contours are the predicted values. It's immediately clear that there is an "island effect" - when smoothed out, ten year economic growth over the North Island of New Zealand has been substantially less than in the South Island.

3.3.2 Methodological strategy

There are well known problems with analysing spatial data as though the observations are independently distributed with regard to their spatial relationship.[2] There are a range of ways that spatial relationships can be taken into account in an economic analysis such as that in this subsection. For example, it is possible to build spatial auto-correlation into the estimation process, so weights are chosen iteratively in response to correlation between spatially adjacent observations. The same estimation methods used for temporal auto-correlation and for mixed effects models can be adapted for this purpose.[1]

One approach that is relatively simple to implement is to include spatial coordinates (eg latitude and longitude of the centres of regions for which observations have been aggregated) as explanatory variables in the model. As relationships between coordinates and the response variable are not expected to be linear, some kind of smoothing and interaction between the two variables is required. Fitting a flexible spline as part of a Generalized Additive Model[24] is an effective way to do this and is the method I choose for this example.

To choose how I allocate degrees of freedom I use the generalised Spearman ρ^2 which is an extension of Spearman's ρ rank correlation. Generalised Spearman ρ^2 is the ordinary R^2 from predicting the rank of Y based on the rank of X and the square of the rank of X.[10] It can detect the strength of nonlinear and nonmonotonic relationships and is an effective way of identifying the importance of explanatory variables that we have already committed to including in the model.

The results of this pre-model-fitting analysis, calculated with the aid of Harrell's `rms` R package,[11] are shown in Figure 14. The relationships between our two economic explanatory variables and ten year growth rates look to be extremely weak. In contrast, the relationship with the spatial variables is strong. We cannot afford more than a single degree of freedom each for our two economic variables, so in our modelling we constrain the relationship between them and economic growth to be linear. We use our four remaining degrees of freedom for a flexible spline based on the interaction of longitude and latitude.

The model can be represented as follows:

$$GDPGrowth_i = \beta_0 + \beta_1 \times PropAg2005 + \beta_2 \times GDPpp2005 + \beta_3 \times s(lat, long, 5) + \epsilon_i$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2)$ and $s(lat, long, 5)$ indicates a smoothing term of dimension 5 for the geographical centres of Territorial Authorities.

Figure 10: Average annual real GDP growth 2005 - 2015

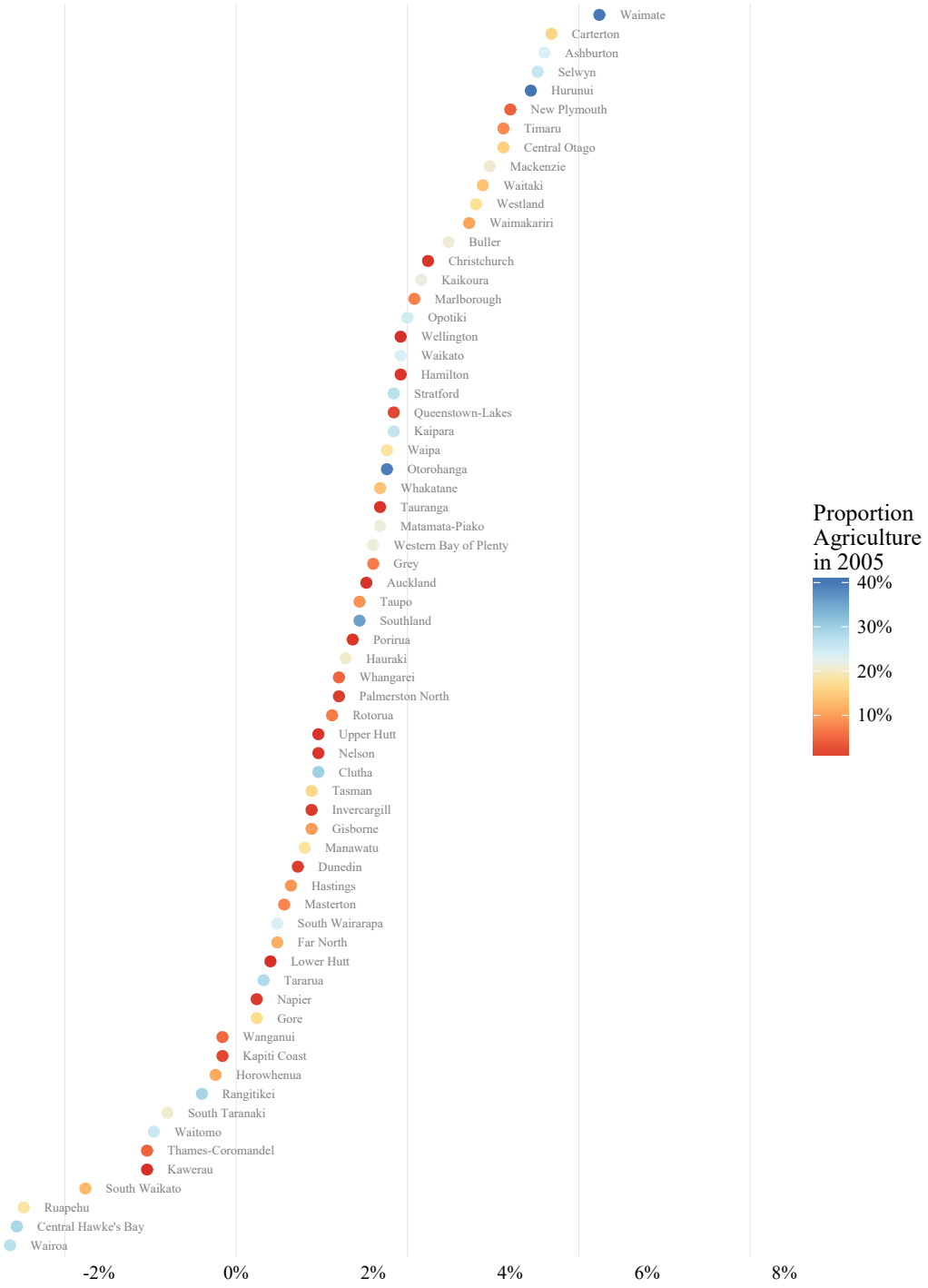


Figure 11: Agricultural focus in 2005 and GDP growth 2005 - 2015

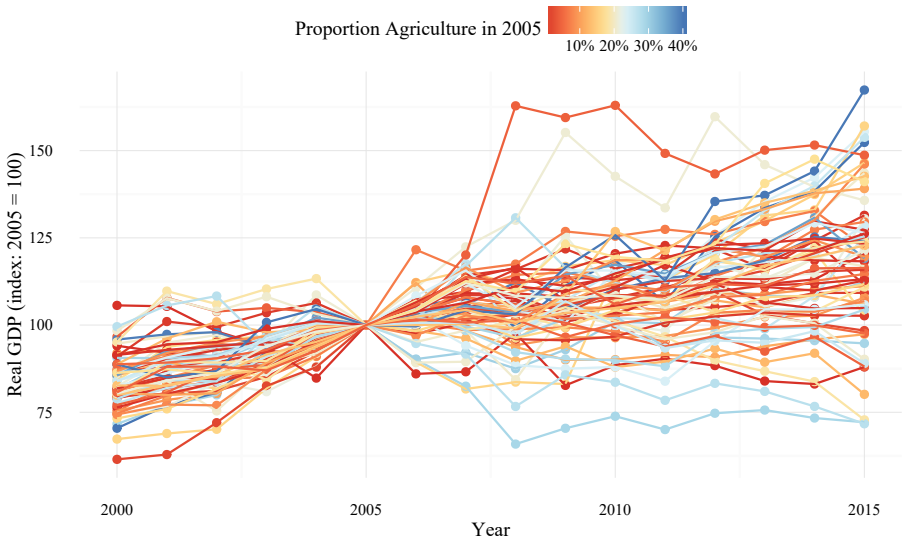


Figure 12: Two economic characteristics in 2005 and GDP growth 2005 - 2015

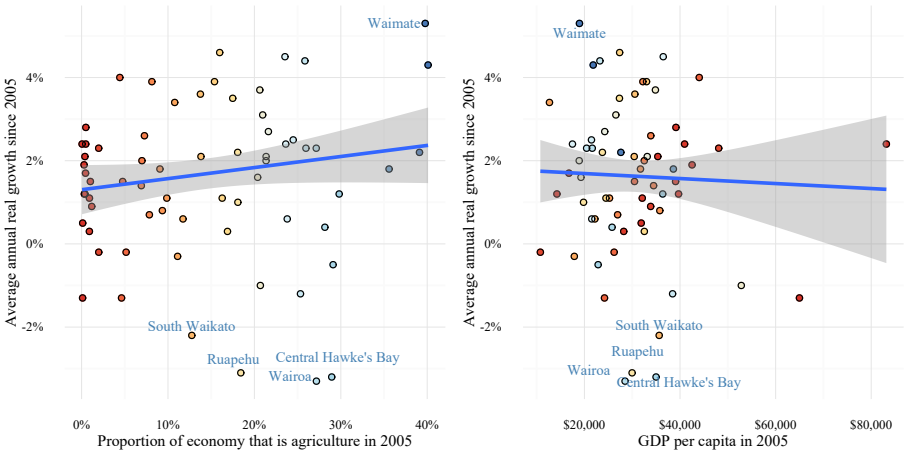


Figure 13: Spatial patterns in GDP growth 2005 - 2015

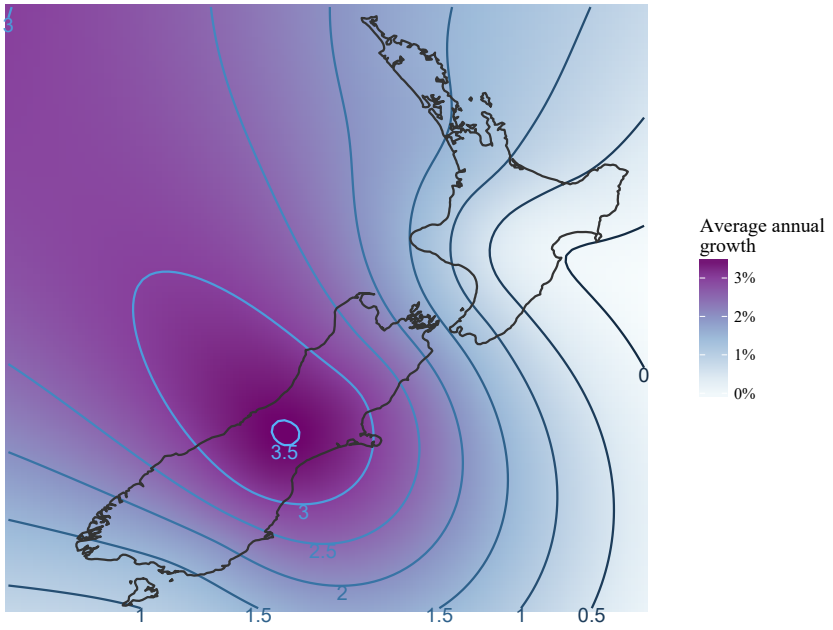
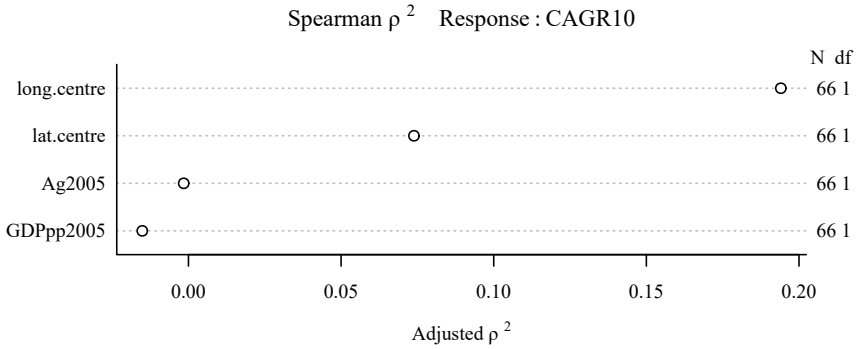


Figure 14: Strength of marginal relationships between predictors and response using generalised Spearman ρ^2



3.3.3 Results

The Generalized Additive Model was fit with the `gam` function from the `mgcv` R package.[24] The results are summarised in Tables 1 and 2. After the visual explorations so far, it is not surprising to see that while there is strong statistical evidence of spatial effect, there is no evidence that either of our two 2005 snapshot economic variables has any relationship with growth from 2005 to 2015.

Table 1: Parametric effects in model of 10 year economic growth

	CAGR10
Ag2005	0.003
	(0.018)
GDPpp2005	−0.000
	(0.000)
Constant	0.018**
	(0.007)
Observations	66
Adjusted R^2	0.317
Log Likelihood	176.856
UBRE	0.000
Notes:	**Significant at the 5 percent level.

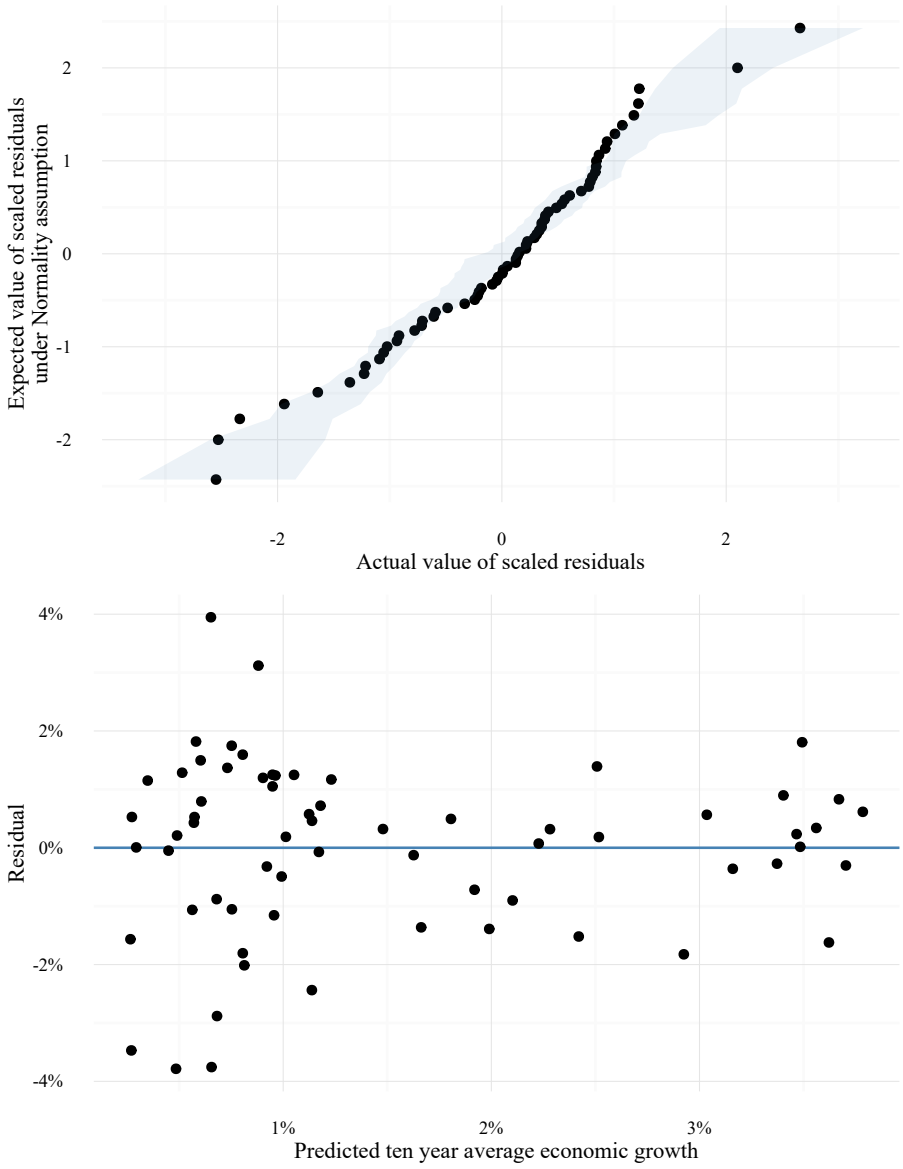
Table 2: Smooth functions in model of 10 year economic growth

	edf	Ref.df	F	p-value
s(long.centre,lat.centre)	3.83	3.98	8.78	0.00

Figure 15 shows two diagnostic plots for the final model. The top graphic compares the actual distribution of residuals to the values they should have under a theoretical Normal distribution, and we see that the error terms are approximately normally distributed as was assumed. The background ribbon in the top graphic shows simulations of 19 genuinely Normally distributed variables as reference. The bottom graphic aims to detect un-modelled shape in the residuals, and similarly can be interpreted as "no obvious problems here".

We can have some confidence that the conclusions from the model are justified on the basis of the evidence before us; and report that we have found no evidence that the 2005 values for GDP per person or industrial focus on agriculture had an impact on subsequent economic growth.

Figure 15: Assumption-checking diagnostic plots for generalized additive model



4 Final comments

4.1 Future work

A range of further work is possible to improve the method of creating these modelled estimates.

- Experiments are under way with data from the LEED aggregated by place of work, rather than place of residence. If this becomes available at the right level of classification, it would replace the need for a "commuting correction" step and materially improve the validity of the estimates.
- If the above improvement isn't possible, the "commuting correction" could still be improved by making use of the 2000 and 2006 censuses (in addition to the 2013 Census currently in use), and possibly by estimation of industry-specific commuting factors.
- The industry-level estimates stop at current year minus three. It might be possible to forecast at the industry level through to current year minus one, possibly by leveraging new data sources not yet integrated into the project.
- A range of existing regional price data could be incorporated to improve the calculation of real GDP by location.

Further work could involve entirely new projects, for example the development of regional expenditure or income estimates to complement these measures of production.

4.2 Accessing the data

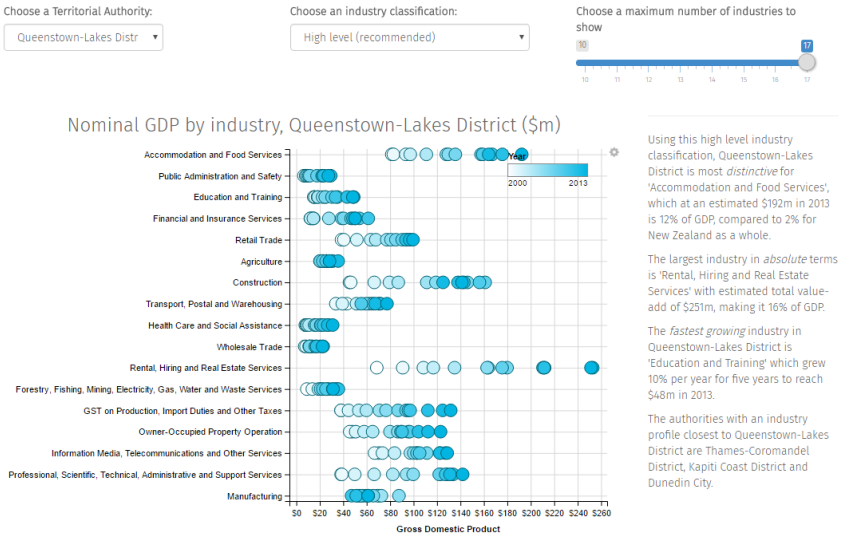
There are currently three main methods of accessing the data.

The primary access point is the web application developed for the purpose using the `Shiny` framework,[4] available at `http://www.mbie.govt.nz/info-services/sectors-industries/regions-cities/research/modelled-territorial-authority-gross-domestic-product/interactive-web-tool`. A screenshot from this web app is provided at Figure 16. It allows the user to easily do all the most commonly expected slices and dices and charts from the data including:

- time series plots of total GDP or selected industry for selected Territorial Authorities and / or New Zealand;
- ability to adjust for population or inflation;
- represent time series as an index rather than in dollar terms, to make it easier to compare across industries and Territorial Authorities;
- show a dot chart of growth rates rather than time series line charts;
- a detailed look at the top industries in a selected region, showing change over time as illustrated in Figure 16;
- machine-generated commentary on the chosen Territorial Authority's distinctive industries, most important and fastest growing industries, and the other Territorial Authorities it most closely resembles;

- scatter plots of industry GDP, comparing two industries with each point representing a Territorial Authority and ability to choose between absolute dollars or proportion of the economy shown on each axis.

Figure 16: Modelled TAGDP web-app - district view



Selected slices of the data have been incorporated into the internet versions of the Regional Economic Activity Report, available at `http://www.mbie.govt.nz/info-services/business/business-growth-agenda/regions`. Both the web-app and the mobile-app feature some of the Modelled Territorial Authority GDP, including GDP and GDP per capita absolute values and growth rates, and selected industry perspectives such as Agriculture as a percentage of GDP. A screenshot is shown in Figure 17.

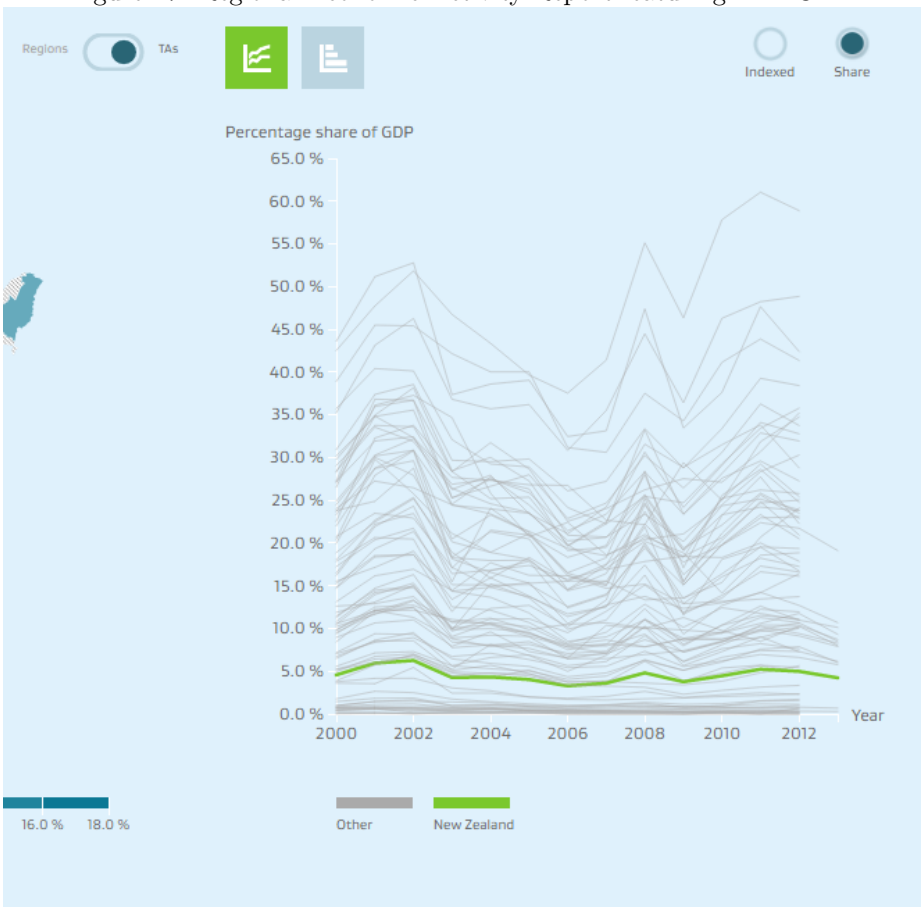
The data can be downloaded in full from `http://www.mbie.govt.nz/info-services/sectors-industries/regions-cities/research/modelled-territorial-authority-gross-domestic-product/data-download` and this is the recommended means of access for analysts seeking to do modelling or reshaping of data in their own tools. Two separate rectangles of data are provided: by industry to 2013; and totals to 2015. Details of the columns in the data are provided at the address referenced above.

The source code that generated the final data from its sources is available as a GitHub repository at `https://github.com/nz-mbie/MTAGDP`.

4.3 Re-use and further analysis

We hope that the data will be re-used extensively. Inevitably areas for improvement will be found in the method or flaws in the details of the

Figure 17: Regional Economic Activity Report featuring MTAGDP



results. Doubtless some interesting findings will turn out to be artefacts of the creation process. However, other interesting findings will provide genuinely new insight into the regional aspects of New Zealand's economy. The only way these issues can be teased out is through open scrutiny of the data and method. MBIE welcomes constructive criticism and suggestions for improvement.

The source code of the analysis in this paper, in R and LaTeX, is available at `https://github.com/ellisp/mtagdp-nzae`.[25221912837]

4.4 Acknowledgements

The development of Modelled Territorial Authority GDP was a team effort from the Sector Trends team in Evidence, Monitoring and Governance Branch of the Ministry of Business, Innovation and Employment. I was the lead designer and developer but Franz Smith made important contri-

butions to the method and to the code implementing it. Twelve months into its life in 2015 when it was languishing and facing possible relegation to the drawer of "projects that never quite made it to production", Dr Smith successfully took over management of the project and has pushed through all three data releases so far. Shaun McGirr, Senay Yasar Saglam, Talosaga Talosaga and James Hogan made crucial contributions of peer review and quality control, both conceptually and in matters of detail. MBIE's web services team have made their usual professional and timely contribution to the actual publication.

The project was helped enormously by several fruitful seminars with NZIER and with the Statistics New Zealand National Accounts team, all of which provided valuable critical and constructive feedback and suggestions. Statistics New Zealand generously provided a custom cut of their Regional GDP data at a lower level of granularity than normally published, which has become a crucial part of the process. Statistics New Zealand also undertook a range of "reality checks" of the results, although MBIE bears all responsibility for the results (see next section).

4.5 Disclaimer

This paper is written as an individual and apart from where explicitly described as otherwise, opinions in it should be attributed to me rather than the Ministry of Business, Innovation and Employment.

Any use of the MTAGDP data should take place in full understanding of the following points:

- These estimates are at a more detailed level of granularity than available in the Statistics New Zealand official Tier 1 regional GDP series. They are experimental in nature and should be used with caution. The data are modelled and produced by the Ministry of Business Innovation and Employment (MBIE) (not by Statistics New Zealand).
- These estimates are not a Tier 1 statistic and have been created by MBIE for research purposes. While various Statistics New Zealand collections form the source data, Statistics New Zealand will not be held accountable for any error, inaccurate findings or interpretation within the data or related publications. One of the sources used for the modelling is a customised dataset created in a way that protects confidentiality, provided by Statistics New Zealand. Access to that data was provided to MBIE by Statistics New Zealand under conditions designed to give effect to the security and confidentiality provisions of the Statistics Act 1975.
- While all care and diligence has been used in processing, analysing, and extracting data and information for this publication, MBIE does not warranty it is error free and will not be liable for any loss or damage suffered by the use directly, or indirectly, of the information.

References

[1] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015.

[2] Roger S. Bivand, Edzer Pebesma, and Virgilio Gomez-Rubio. *Applied spatial data analysis with R, Second edition*. Springer, NY, 2013.

[3] Winston Chang. *extrafont: Tools for using fonts*, 2014. R package version 0.17.

[4] Winston Chang, Joe Cheng, JJ Allaire, Yihui Xie, and Jonathan McPherson. *shiny: Web Application Framework for R*, 2016. R package version 0.13.2.

[5] Gabor Csardi and Tamas Nepusz. The igraph software package for complex network research. *InterJournal*, Complex Systems:1695, 2006.

[6] Peter Ellis. *Network charts of commuting in New Zealand with R and D3*, 2015. `http://ellisp.github.io/blog/2015/12/26/commuting-network`.

[7] Peter Ellis. *ggseas: 'stats' for Seasonal Adjustment on the Fly with 'ggplot2'*, 2016. R package version 0.4.0.

[8] Toby Dylan Hocking. *directlabels: Direct Labels for Multicolor Plots*, 2015. R package version 2015.12.16.

[9] Rob J Hyndman, Earo Wang, Alan Lee, and Shanika Wickramasuriya. *hts: Hierarchical and Grouped Time Series*, 2016. R package version 5.0.

[10] Frank E. Harrell Jr. *Regression Modeling Strategies*. Springer, New York, 2001. ISBN 0-387-95232-2.

[11] Frank E Harrell Jr. *rms: Regression Modeling Strategies*, 2016. R package version 4.5-0.

[12] David Kahle and Hadley Wickham. ggmap: Spatial visualization with 'ggplot2'. *The R Journal*, 5(1):144–161, 2013.

[13] L. Kaufman and P. J. Rousseeuw. *Finding groups in data: an introduction to cluster analysis*. John Wiley and Sons, New York, 1990.

[14] Thomas Lumley. *survey: analysis of complex survey samples*, 2014. R package version 3.30.

[15] Martin Maechler, Peter Rousseeuw, Anja Struyf, Mia Hubert, and Kurt Hornik. *cluster: Cluster Analysis Basics and Extensions*, 2016. R package version 2.0.4.

[16] Ministry of Business Innovation and Employment New Zealand. *Modelled Territorial Authority Gross Domestic Product: Experimental estimates to help research leverage New Zealand's official statistics - summary document*, 2015. `http://www.mbie.govt.nz/info-services/sectors-industries/regions-cities/research/modelled-territorial-authority-gross-domestic-product`.

[17] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016.

[18] Christoph Sax and Peter Steiner. *tempdisagg: Methods for Temporal Disaggregation and Interpolation of Time Series*, 2014. R package version 0.24.0.

[19] Kamil Slowikowski. *ggrepel: Repulsive Text and Label Geoms for 'ggplot2'*, 2016. R package version 0.5.

[20] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002. ISBN 0-387-95457-0.

[21] January Weiner. *riverplot: Sankey or Ribbon Plots*, 2015. R package version 0.5.

[22] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2009.

[23] Hadley Wickham and Romain Francois. *dplyr: A Grammar of Data Manipulation*, 2015. R package version 0.4.3.

[24] S.N. Wood. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B) 73(1):3-36*, 2011.

[25] Yihui Xie. *knitr: A General-Purpose Package for Dynamic Report Generation in R*, 2016. R package version 1.13.

[26] Statistics New Zealand. *Estimated Resident Population for Territorial Authority Areas, at 30 June(1996+) (Annual-Jun)*, 2016.

[27] Statistics New Zealand. *Linked Employer Employee Data*, 2016.

[28] Statistics New Zealand. *NZ Business Demography Statistics*, 2016.

[29] Statistics New Zealand. *Regional Gross Domestic Product*, 2016.

[30] Statistics New Zealand. *SNE – Series, GDP(P), Chain volume, Actual, ANZSIC06 industry groups (Annual-Mar)*, 2016.

[31] Statistics New Zealand. *SNE – Series, GDP(P), Nominal, Actual, ANZSIC06 industry groups (Annual-Mar)*, 2016.