MINISTRY OF BUSINESS,
INNOVATION & EMPLOYMENT
HĪKINA WHAKATUTUKI

Estimates of GDP by detailed industry and Territorial Authority

Experimental estimates to help research leverage New Zealand's official statistics - Methodology Document

June 1, 2016

New Zealand Government

MINISTRY OF BUSINESS, INNOVATION & EMPLOYMENT
HĪKINA WHAKATUTUKI

Ministry of Business, Innovation and Employment (MBIE)

Hīkina Whakatutuki – Lifting to make successful

MBIE develops and delivers policy, services, advice and regulation to support economic growth and the prosperity and well-being of New Zealanders.

MBIE combines the former Ministries of Economic Development, Science + Innovation, and the Departments of Labour and Building and Housing.

More information

www.mbie.govt.nz

0800 20 90 20

Information, examples and answers to your questions about the topics covered here can be found on our website www.mbie.govt.nz or by calling us free on 0800 20 90 20.

Disclaimer
This document is a guide only. It should not be used as a substitute for legislation or legal advice. The Ministry of Business, Innovation and Employment is not responsible for the results of any actions taken on the basis of information in this document, or for any errors or omissions.

June 1, 2016

Contents

List of Figures

List of Tables

1. Summary

From October 2015, the Ministry of Business, Innovation and Employment will annually publish experimental estimates of Modelled Territorial Authority Gross Domestic Product (MTAGDP[1]). The rationale is to create a public asset to promote further research and analysis on industry production trends with a regional focus. Ultimately, the aim is to better inform decision-making at the various levels of government and by private decision-makers (e.g. investors). The estimates are created from publicly available data and the source code used to create them is available for inspection and comment. This document serves to describe the methodology of deriving the MTAGDP estimates and a 'users guide' for the source code that was used to generate the estimates.

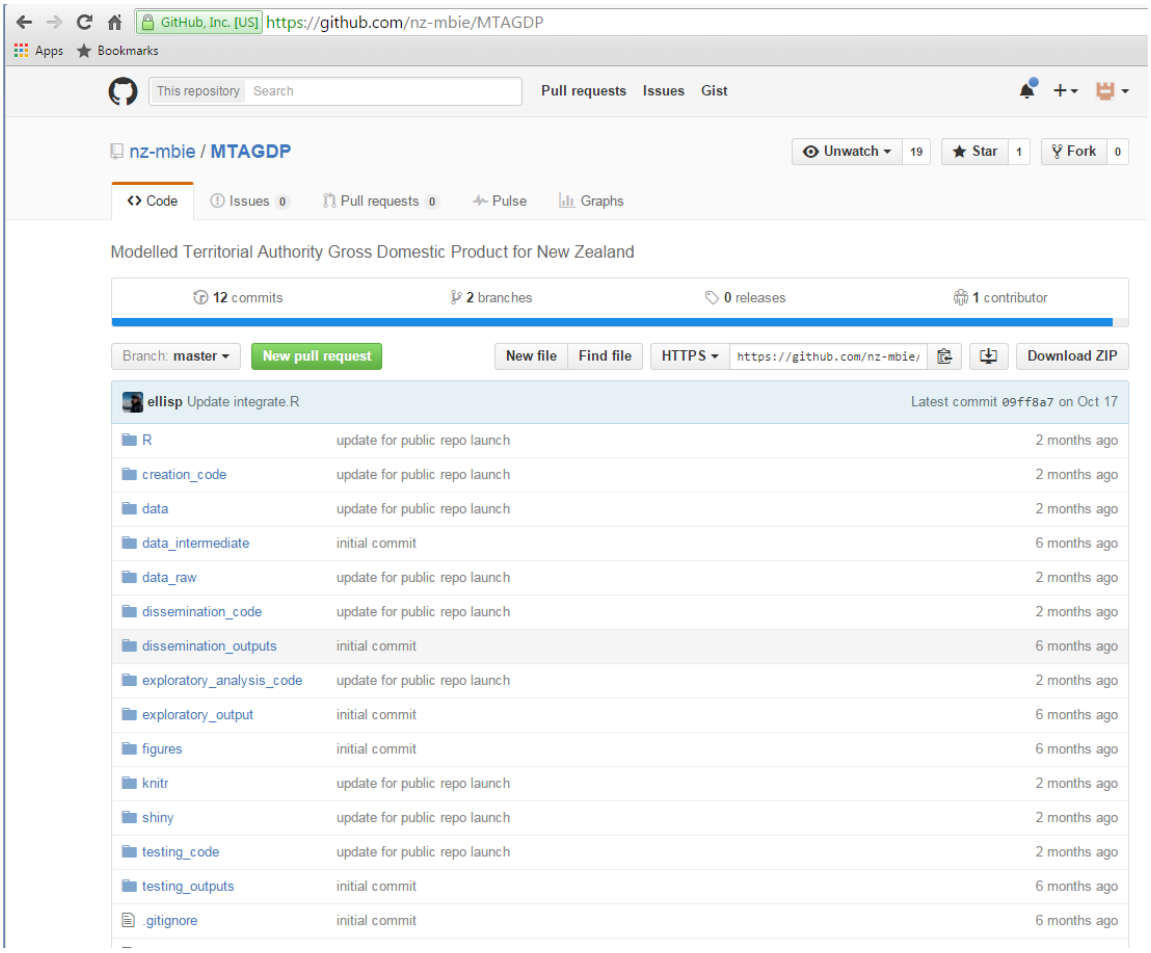
The methodology includes:

- Overview of the statistical methods used to generate the MTAGDP estimates and key assumptions of the technique for this application.
- Description of the code structure and work flow for calculating MTAGDP, including the import of raw data, concordances for linking industries and geographic regions
- Routines for testing the output against official published statistics
- Code to generate graphic outputs of main results
- General notes on the limitations and caveats

Outputs from this project will be featured in the Regional Economic Activity Report and its accompanying web and mobile apps from October 2015 onwards. MBIE has identified a range of possible improvements that may be implementable in 2016 and onwards. In the meantime feedback is welcomed.

[1]Caveats and disclaimers: 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), according to the methods outlined in this document. 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 publication. 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 gives 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.

Figure 1: Screenshot of the Github repository showing the basic folder structure of the MTAGDP project



2. Introduction

2.1 Who this document is for

This document provides an overview of the methodology and source code used to generate the *Modelled Territorial Authority Gross Domestic Product (MTAGDP)* experimental estimates and several of the graphic outputs. The audience for this document is for persons interested in more background of the methodology, advanced users of the data, and researchers.

This document:

- Provides details of the statistical methods (i.e. Iterative Proportional Fitting, also known as 'raking') and key assumptions for this application.
- Describes the code structure and work flow of the project, including the import of raw data, concordances for linking industries and geographic regions
- Describes tests used to validate the output and interpretations
- Describes the code used to generate graphic outputs for the main results
- Discusses limitations and caveats

2.2 Context

Although there are several sources of economic indicators for the regions of New Zealand, the official published statistics for Gross Domestic Product from Statistics New Zealand are only available to the Regional Council level, with a coarse level of industry distinction (i.e. 4 digit Australian New Zealand Standard Industry Classification (ANZSIC06)). For some research and policy development, there is a need to have a finer level of geographic and industry information (e.g. comparing 'urban' and 'rural' districts within a region). However, alternative sources for this data from consultancy firms and think-tanks are sold as a service, preventing:

- Open scrutiny of the underlying methodology and assumptions as part of an open scientific process
- Reproducibility of results and the ability to incorporate enhancements made by other analysts, leading to a increasingly fragmented source of information for understanding sub-regional economies.

Further, MBIE's understanding is that some of these data do not reconcile to official published Regional or National GDP figures from Statistics New Zealand (this is not possible to confirm, for the reasons noted in the above paragraph). For end-users of this information, this means that inter-regional or inter-annual comparisons are less likely to be reliable and may contradict Tier 1 statistics when aggregated.

2.3 Rationale for making the source code available

As none of the existing sources of regional economic information provided a solution to the fundamental need for access to granular regional and industry-level economic data, MBIE's Sector Trends team initiated this project with the following aims:

- To derive publishable estimates of GDP at the Territorial Authority level and industry classifications equivalent to those used for national GDP
- Establish a web-based tool which allows non-specialist users to explore the data and generate basic analytical graphics.
- Make the data available as a resource for researchers and policy development
- Make the source code and data available to promote transparency and future improvements in the approach

As the private sector already operates in this area (providing information as a service), the decision to provide freely available data was taken only after serious consideration. The benefits of publishing the data in its raw format are expected to include reproducibility and scrutiny leading to improvements in the estimates over time; and consultancy and analytical projects can commence with a head start (compared to previously) rather than allocating resource to repeatedly purchasing data.

Making the source code available provides researchers and advanced users of the data better understand and scrutinise the underlying methodology and assumptions made in the production of the raw data. Additionally, there is an added advantage in providing the ability to reproduce the MTAGDP estimates. Making the source code available also provides a means for contributions for improving the code, the numerical methods, and testing framework, which are likely to lead to improved MTAGDP estimates.

As the current 'snapshot' of the project available on Github relies on particular data infrastructure available within MBIE, the public version will not successfully run to completion. In future releases, MBIE aims to create a version which has the same functionality as our internal working version. At present, the code is provided to illustrate the methods and documentation of how MTAGDP estimates were derived.

3. Methodology

3.1 Overall approach

MBIE has chosen to use "total earnings" published by Statistics New Zealand from the Linked Employer - Employee Database (LEED) as the basis for allocating GDP to a level of granularity below that possible for the Tier 1 statistics. In addition, employee numbers in the Business Demography Statistics (BDS) were also used for fine grained allocations within industries when earnings were not available at the necessary level of a granularity.

The basic approach for deriving estimates of MTAGDP uses a technique called Iterative Proportional Fitting, more commonly known as raking, which adjusts a table of cells to add up to the marginal totals that are known to be accurate. This methodology is commonly used in survey sampling, where data from a sample are raked to provide a set of weights that match the known population totals in categories like age group, sex, and ethnicity.

For estimating MTAGDP, several data sources were used to successively rake to align totals from finer to coarser-level information of geographic regions and industry classifications. In effect, the employee and earnings numbers are weighted up to match the published GDP totals, providing a robust method for estimating GDP. This is a viable approach because ratios of earnings to GDP in different combinations of regions and industries can be determined from the published data in national GDP (i.e. NGDP), regional GDP (i.e. RGDP), and the Linked Employer Employee Database (i.e. LEED).

Using this approach, MBIE's MTAGDP estimates add up to the Gross Domestic Product Tier 1 official statistics published by Statistics New Zealand at the national and Regional Council level (with varying degrees of industry detail). Other than rounding error, MTAGDP matches all currently published results whenever it is aggregated to the necessary level.

An alternative way of thinking about the MTAGDP method is to consider it as taking the official published totals of GDP and using more granular data sources (also published by Statistics New Zealand) to allocate out that GDP to further levels of detail. The methodology models proportions that best fit the known relations of GDP to those other variables.

3.2 Main data sources

An overview of the data sources and the level of detail associated with each included:

Data set	Description
Business Demography Statistics	Yearly two-way table of employee numbers, Territorial Authority by fine level (i.e. 6 digit ANZSIC). http://nzdotstat.stats.govt.nz
Linked Employer-Employee Database Table 4	1-Way: Earnings by Quarter and fine level (i.e. 6 digit ANZSIC06) industry classification. http://nzdotstat.stats.govt.nz
Linked Employer-Employee Database Table 37	1-Way: Earnings by Quarter and Territorial Authority. http://nzdotstat.stats.govt.nz
Linked Employer-Employee Database Table 18	2-Way: Earnings by Quarter and Region, and medium level industry (i.e. 3 digit ANZSIC). http://nzdotstat.stats.govt.nz
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). Statistics New Zealand
National Gross Domestic Product Production Measure, nominal (NGDP)	Provides more detailed industry breakdown than NGDP but no regional information. http://www.stats.govt.nz/infoshare
Regional Gross Domestic Product (RGDP)	Provides regional GDP by industry up to 2012 (at the time of writing). National totals match (within rounding error) those in NGDP. http://www.stats.govt.nz/infoshare

3.3 Techniques

The sequence of raking begins with the Business Demography Statistics (BDS), followed by the LEED tables. This transformed the original granularity (i.e. in the finest available form, but a poor measure of production (i.e. employee numbers from BDS)) to a table with the spatial granularity (i.e. Territorial Authority) and better measure of production (i.e. earnings from the LEED tables). The final stage of raking scaled the earnings values to be proportional to the Regional and National GDP figures - which created estimates of the correct production measure (i.e. GDP) at the desired spatial granularity (i.e. Territorial Authority).

The approach was to use the BDS for a fine level detail for the number of employees in each detailed industry in each Territorial Authority. “Employees” do not include the self-employed, and employee counts do not contain information on pay levels or profits. However on plausible assumptions the information contained in these job numbers can be of great value. The method employed a two step process to develop estimates of gross domestic product at the same level of granularity as the original BDS data:

1. Scale the job numbers to match earnings levels
2. Scale the resulting earnings estimates to match gross domestic product

In step 1, the BDS job numbers were scaled up to the marginal total earnings from the three LEED sources. This required iterative proportionate fitting to three different sets of marginal totals. This can be easiest done with a simple trick: treating the original BDS data as though it were a sample survey, with the number of Employees sample sizes and the marginal Earnings totals from the LEED as a population to weight up those samples to match. The operative part of the R code that does this is reproduced below (hopefully naming conventions are self-explanatory):

```
  # create earnings object from BDS data
    EarningsDetailed <- BDS

  # This is treated as a survey design, with the estimates of employees as weights
    EarningsDetailed_svy <- svydesign(~1, data=EarningsDetailed, weights = ~Employees)

#----------------Weight up to LEED total earnings--------------------

  for(i in 1:5){
     EarningsDetailed_svy <- rake(EarningsDetailed_svy,
                                  sample.margins = list (~Year + LEED18Industry + LEED18Region),
                                    population.margins = list(leed18_pop),
                                    control = list(maxit = 25, epsilon = 100, verbose=FALSE))

     EarningsDetailed_svy  <- rake(EarningsDetailed_svy,
                                    sample.margins = list (~Year + TA_Region_modified),
                                    population.margins = list(leed37_pop),
                                    control = list(maxit = 25, epsilon = 100, verbose=FALSE))

     EarningsDetailed_svy <- rake(EarningsDetailed_svy,
                                   sample.margins = list (~Year + LEED4Industry),
                                   population.margins = list(leed4_pop),
                                   control = list(maxit = 25, epsilon = 100, verbose=FALSE))
  }
```

In step 2, the resulting estimates of Earnings were scaled up to GDP to match the marginal totals of regional GDP (Region by Industry) and national GDP (more detailed industry), using the same technique. Note that in the code below, the data frame `TAGDP` is almost identical to the data frame `EarningsDetailed` created above; the difference arises from complications with time period and with owner-occupied dwelling operations, detailed in a subsequent section of this paper.

```
 # first create a survey object using the commuting_corrected earnings as the weights
   TAGDP_svy <- svydesign(~1, data=TAGDP, weights = ~Earnings_commuting_corrected)

 #----------------------Rake to the various GDP measures----------------------
 # This step iterates manually rgdp v. ngdp to help with harmonising the weights and isolating
 #    any errors.

   for(i in 1:15){
     cat("Beginning of loop\n")
     TAGDP_svy <- rake(TAGDP_svy,
        sample.margins     = list (~Year + RGDPIndustry_custom + RGDP_Region),
      population.margins = list(rgdp_pop_custom[!is.na(rgdp_pop_custom$RGDPIndustry_custom),
```

```
                                c("Year", "RGDPIndustry_custom", "RGDP_Region", "Freq")]),
    control            = list(maxit = 100, epsilon = tolerance, verbose=FALSE))

  TAGDP$GDP <- weights(TAGDP_svy)

  # First to the national gdp published industry figures
    TAGDP_svy <- rake(TAGDP_svy,
                  sample.margins = list (~Year + NGDP_industry),
                  population.margins = list(ngdp_pop),
                  control = list(maxit = 100, epsilon = tolerance, verbose=FALSE))
    TAGDP$GDP <- weights(TAGDP_svy)

  # Then to the two iterations of RGDP.  All these should be very consistent
    TAGDP_svy <- rake(TAGDP_svy,
                  sample.margins = list (~Year + RGDP_Region + RegionIndustryRGDP15),
                  population.margins = list(rgdp_pop_pub_det),
                  control = list(maxit = 100, epsilon = tolerance, verbose=FALSE))
    TAGDP$GDP <- weights(TAGDP_svy)

     TAGDP_svy <- rake(TAGDP_svy,
                   sample.margins = list (~Year + RGDP_Region + RGDP_industry),
                   population.margins = list(rgdp_pop_pub),
                  control = list(maxit = 100, epsilon = tolerance, verbose=FALSE))
  # how are we going: troubleshooting
    cat(sum((TAGDP$GDP - weights(TAGDP_svy)) ^ 2),
                          "change after weighting to RGDP high level on loop ", i, "\n")
   TAGDP$GDP <- weights(TAGDP_svy)

}
```

As the custom Regional GDP data provided by Statistics New Zealand included a number of 'confidentialised' cells within the data table, the missing values were imputed using the IPF (or 'raking') process described above. Here, the marginal totals from the published Regional and National GDP industries were used to estimate the values of the missing cells.

Because the method relies on the detailed industry data provided in the Business Demography Statistics and published Regional GDP figures, the MTAGDP estimates are available from the year 2000 to 2012. As the detailed industry data are released every three years, MTAGDP will be updated after the release Regional and National GDP figures in March 2016.

3.4 Incorporating Commuter Information

One limitation of using different employer-employee sources for deriving estimates of GDP at the TA level was the BDS provides information of number of employees at the business address, while 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 (e.g. 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 transferred

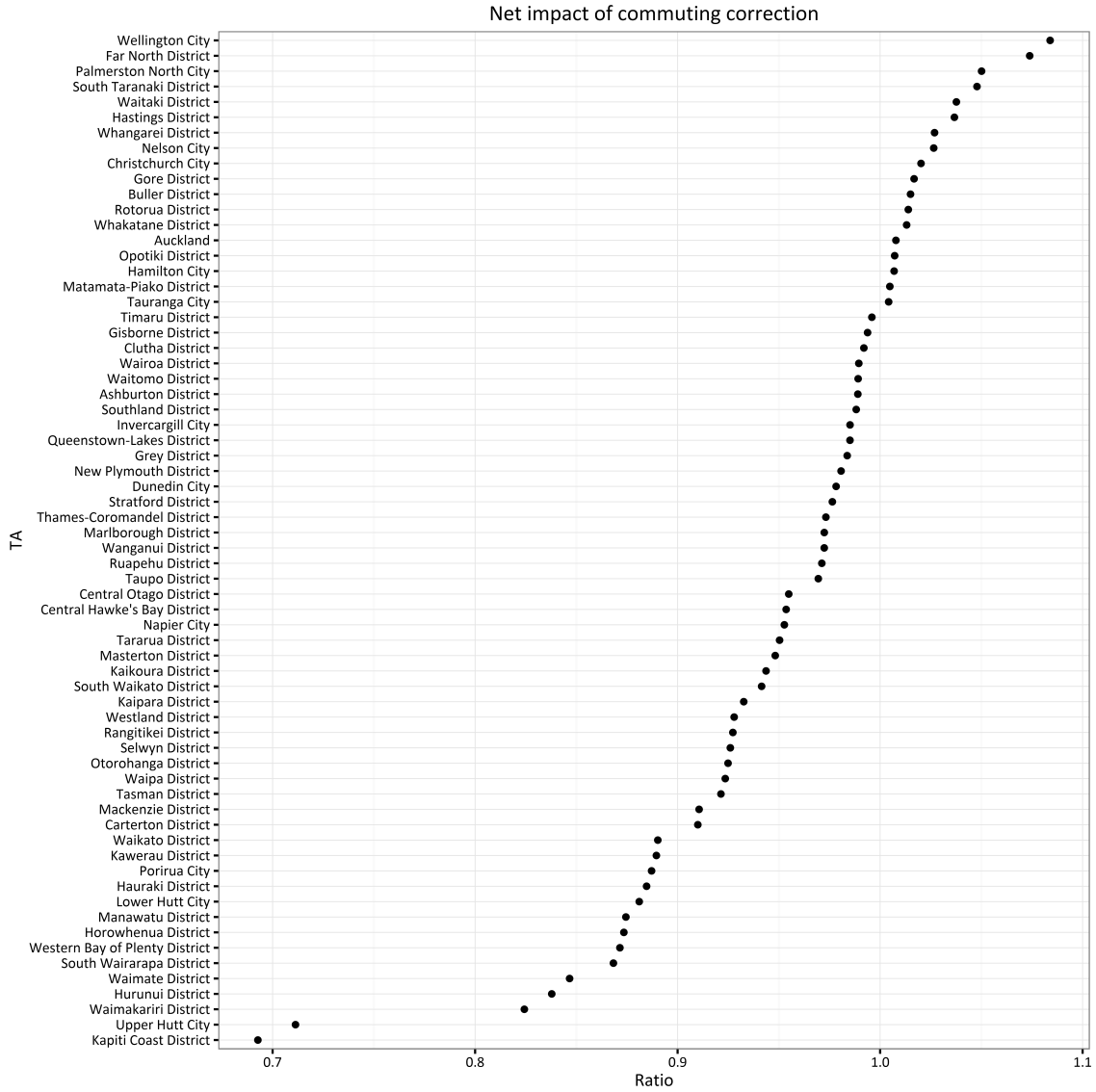
across Territorial Authority boundaries.

In order to correct for the transfer of earnings across Territorial Authority boundaries, data from workplace and home addresses from the 2013 Census[2] were used to calculate the relative proportion of earnings based on the reported commuter numbers. The four largest destination 'commuting' Territorial Authorities were used to characterise the majority of commuting flows from 'origin' TAs. Although this approach does not include industries and occupation classes that has different proportion of commuters, it provides a reasonable approximation with available data. Improving this has been noted as an area for future improvement.

Analysis of the commuter flows using the ratio of earnings between the 'corrected' values and the original values illustrate regions that there are about 20 Territorial Authorities that receive a greater number of commuters (i.e. having a ratio greater than 1) relative to 'donor' TAs (Figure 2).

[2] http://www.stats.govt.nz/Census/2013-census/profile-and-summary-reports/commuter-view-visualisation.aspx

Figure 2: Net impact of commuting correction



3.5 Inflation-adjusted and per capita measures

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's 'SNE - Series, GDP(P), Nominal, Actual, ANZSIC06 industry groups (Annual-Mar)' and 'SNE - Series, GDP(P), Chain volume, Actual, ANZSIC06 industry groups (Annual-Mar)'[3]. This provided industry-level

[3] http://www.stats.govt.nz/infoshare

deflators for 31 industry categories.

The MTAGDP estimates also include per capita estimates for GDP. Population totals by TA came from Statistics New Zealand's Estimated Resident Population for Territorial Authority Areas, at 30 June(1996) (Annual-Jun)[4].

3.6 GDP Estimates for 'between years' of Business Demography Statistics and LEED tables

Policy development and research requires access to up-to-date data. As part of the source data for creating MTAGDP (i.e. Business Demography Statistics, Linked Employer–Employee Database) are only available every three years, the last year of the present MTAGDP series is 2012. In an effort to support policy development and researchers, a method was developed to make MTAGDP estimates for 'between' years of publishing BDS and LEED tables. The aim of the forecasting was to extend the MTAGDP series to match annual published total Regional and National GDP figures from Statistics New Zealand.

The method used for extending the MTAGDP series is based on traditional time-series forecasting techniques. A technique referred to as *Hierarchical [or Grouped] Time Series* uses disaggregated relationships of distinct categories in a time series (e.g. product types in geographic locations) in a way that reconciles across levels in the hierarchy or across groups [5]. In other words, these methods allow for forecasts at each level to be summed to provide information of the level above or across groups.

The MTAGDP data series is well suited for this technique, as the modelled data represent industries within Territorial Authorities as natural grouped sets. And, because the forecast estimates can be aggregated into Regional and National GDP figures, it provides a means to ensure that the forecasted values match the published total values at these levels.

Summaries of total GDP at the TA level were used to forecast GDP estimates out 2 years to match the last published Regional and National GDP figures published by Statistics New Zealand. As the data series provided few time steps for time series forecasting (i.e. yearly data from 2000 to present), MTAGDP data was disaggregated into monthly time steps to provide a greater number of points for developing an ARIMA model[6]. After forecasting, these data were aggregated back to the original yearly time steps.

The grouped time series function used an 'optimal combination' forecasting method, which creates an independent forecast for each series in the hierarchy. As these forecasts are created independently, they do not necessarily add up according to the hierarchy or group structure. The 'optimal combination' method combines the independent forecasts and generates a set of revised

[4] `http://www.stats.govt.nz/infoshare`

[5] Hyndman, R.J., R.A. Ahmed, G. Athanasopoulos and H.L. Shang (2011) Optimal combination forecasts for hierarchical time series. Computational Statistics and Data Analysis, 55(9), 2579–2589. `http://robjhyndman.com/papers/hierarchical/` `http://robjhyndman.com/hyndsight/gts/` `https://cran.r-project.org/web/packages/hts/index.html`

[6] Using the R package *tempdisagg* `https://cran.r-project.org/web/packages/tempdisagg/index.html`

forecasts that resemble the univariate forecasts but also aggregate consistently with the hierarchical or group structure.

As a final step in the modelling, published data from Regional and National GDP figures were used to adjust the forecast results to coincide with official public statistics. This used the same methodology of 'iterative proportional fitting' (otherwise known as 'raking') used to create the original MTAGDP series, which makes the series internally consistent and matches published values for the 'between years'.

GDP estimates were also converted into 'real' and 'per capita' measures. 'Real GDP' measures were created by summarising the 2000-2012 'real_GDP' values and adjusting the 2013 and 2014 estimates according to the "Rolling Annuals - IPDs, Actual, Total (Annual-Mar)" Infoshare table. Values were deflated according to the 2012 values to be consistent across the entire series. Population data at the TA level were taken from the "Estimated Resident Population for Territorial Authority Areas, at 30 June(1996+) (Annual-Jun)" table in Infoshare.

Final output was compiled as a separate series into a single `.csv` file to accompany the MTAGDP file with the industry-level information. Future incorporation of these data into the shinyapp is planned for 2016.

4. Workflow & Code Structure

4.1 Workflow

The project was designed to compile the data inputs necessary for computing MTAGDP, perform the calculations, test outputs, generate dissemination graphics, and save final data objects. This process is handled through executing the `integrate.R`, which performs the following sequence:

- Loads the necessary packages and ancillary functions
- Imports concordances, raw data, and additional data sources
- Performs adjustments to "harmonise" across tables of similar data (e.g. employment, GDP)
- Creates `survey` objects for raking to calculate MTAGDP
- Saves the resulting data object & outputs `.csv` file
- Tests and trouble-shoots results
- Creates graphic outputs of the data

Comments within the `integrate.R` script provides a short-form description of each step of the compilation process, while the header of individual scripts sourced by `integrate.R` provide additional detail particular to that particular function or analysis. Additional comments in the body of individual scripts also provide documentation of particular steps. In general, these are kept to a minimum.

4.2 Folders and File Descriptions

The project structure contains a number of folders for holding different data types, scripts, and source code functions. The folder `creation_code` contains scripts that import the data, make adjustments to ensure they are consistent across different tables, calculate MTAGDP and save the data, including:

File in the 'creation_code' Folder	Explanation
import_concordances.R	Imports tables to align Territorial Authorities to Statistics NZ LEED regions, and industry classes for published National and Regional GDP.
create_commuting_corrections.R	Imports 2013 Census table of usual residence by workplace address table and calculates the 4 largest commuter flows across Territorial Authorities, and saves the output as individual `.csv` files. Script also defines a function used to adjust earnings across Territorial Authorities prior to raking.

import_gdp_deflator_totals.R	Imports Nominal and Chain volumes for National GDP by ANZSIC06 industry groups to calculate industry-level deflators. Used in the calculation of real GDP.
import_population_totals.R	Imports estimated resident population numbers by Territorial Authority, filtered for the time series of the MTAGDP estimates (i.e. 2000 to 2012). Used in the calculation of per-capita measures of GDP for industries.
import_BDS.R	Imports Business Demography Statistics (BDS) of employee counts by industry and Statistics NZ regions. Data are concorded for Territorial Authority and industry classes. Used in the first iteration of raking to match employee numbers with LEED tables.
import_leed4.R	Imports data from LEED table 4 for Total Earnings, all Industries and Year quarters from 2000 to 2012. Used in the raking to employee numbers from the BDS.
import_leed18.R	Imports data from LEED table 18 for ANZSIC06 industry classes and regions. Used in the raking to employee numbers from the BDS.
import_leed37.R	Imports data from LEED table 37 for earnings by Territorial Authority. Used in the raking to employee numbers from the BDS.
import_leed37.R	Imports data from LEED table 37 for earnings by Territorial Authority. Used in the raking to employee numbers from the BDS.
import_RGDP.R	Imports Regional GDP data from different sources, including Statistics NZ public version of RGDP and a custom version of 30 industries x 15 regions. Two versions of the public version are created, one including modified Region x industry categories to aid in concording to the `data_raw/concordances/industries.csv` file. Used in the 2nd iteration of raking to match the marginal totals of earnings (adjusted for commuting across TAs).
impute_rgdp_custom.R	Script imputes the "confidentialised" cells in the custom version of Regional GDP provided by Statistics NZ. This uses the `rake` function from the `survey` package to fill missing values and re-scale the final data object. Final data object used in the 2nd iteration of raking in `grunt.r`.

import_NGDP.R	Imports nominal, actual National GDP vales for detailed ANZSIC06 industry classes. Data are modified to include GST and Duties Tax from the National total table published by Statistics NZ.
harmonise_leed_totals.R	Comparing the LEED table totals has some inconstant marginal totals, so these are adjusted to match the LEED 37 earnings totals.
harmonise_GDP_totals.R	This script adjusts the marginal totals of the different versions of GDP for Regional and National levels.
grunt.R	This script calculates Territorial Authority GDP through different iterations of raking, converting the employee numbers to earnings and earnings to GDP while retaining the finest level of detail at industry levels (ANZSIC06) and Territorial Authorities.
modify_tagdp.R	This script modifies the resultant data object from `grunt.R` by adjusting for inflation (i.e. for real GDP), calculating per capita measures, and combining industry levels to match published National and Regional GDP. Data are also summarised for Regional Councils and Territorial Authorities.

The `data_raw` folder contains raw data downloaded from Statistics New Zealand's InfoShare and nzdotstats sites, `.csv` files from data queries from MBIE's internal data bases (reflecting InfoShare tables), and custom `.xlsx` data files. There are also two additional data folders (i.e. `data_intermediate` and `data` for storing intermediate and final data files. A sub-folder contained within includes the concordances for matching geographic regions and industry codes, including:

File in the 'data_raw/concordances' Folder	Explanation
industries.csv	This file defines the concordances between the different levels of industry classification from the LEED tables and Regional and National GDP data sources.
leedTA_to_SNZTA.csv	This file serves to link the Territorial Authorities published in the LEED tables and the standard TA definitions.
region_to_leed18_andRGDP_region.csv	This file links the Regions published in the LEED18 tables and published Regional GDP data.
RegionIndustryRGDP15.csv	This file concords the industry levels between the detailed Regional GDP published in March 2015 and previous RGDP versions.
TA_to_multiple_regions.csv	This files links the Statistics NZ Territorial Authorities and Regional Councils and modified TAs where the TA and RC boundaries do not align. Relative proportions allocated to the Regional Councils are also defined.
deflatorIndustries.csv	This file serves to link the LEED4 industries with industries from the deflator tables.

In the calculation of MTAGDP, a number of tests are performed for checking the matching of marginal totals of the different data sources for earnings (i.e. from the LEED tables) and GDP. Some of the testing prints results to the screen, while others provide plots that are saved to the `testing_outputs` folder. The key testing files are detailed below:

File in the 'testing_code' Folder	Description
compare_marginal_totals.R	Script compares marginal totals across the different data sources for National and Regional GDP, and LEED tables. Comparisons are printed to the screen and summarised in figures combined into a single `.pdf` file.
troubleshoot_incompatible_population.R	Script makes comparisons between marginal totals from the different data sources for Regional and National GDP and LEED tables. Individual tests are used to validate sources used in the stages of calculating MTAGDP.
net_impact_commuting_correction.R	Provides a visual summary of influence of commuter correction on earnings. The expectation is that dormitory suburbs will have ratios less than 1 and urban centres with large numbers of commuters greater than 1.
testing_wellington_ratios.r	Provides a visual assessment of the ratio of GDP to earnings from the LEED tables after the 'commuting correction' is applied. Figure is saved to a `.pdf` file.

The folder `dissemination_code` provides scripts which produce figures and additional analyses of the resultant MTAGDP estimates. These scripts include the data preparation necessary for the deployment of the web visualisation tool (via shinyapps). Lastly, there is a script which produces a snapshot of the current project for public release to Github. These files include:

File in the 'dissemination_code' Folder	Explanation
plot_commuting_patterns.r	Creates a network diagram for the commuting data from the 2013 Census used in correcting the earnings data prior to raking.
plot_industries_by_region.R	Creates figures illustrating the top industries in GDP terms for each Regional Council. Plots are combined into a single `.pdf` file.
plot_industries_by_TA.R	Creates figures illustrating the top industries within Territorial Authorities in GDP terms. Plots are combined into a single `.pdf` for public release.
plot_topTAs_by_industry	Creates figures illustrating the Territorial Authority rankings for individual industries (i.e. for National GDP industry classes) in GDP terms. Plots are combined into a single `.pdf` file for public release.
prepare_basic_TA_commentary.R	Generates commentary for the shinyapp, including most distinctive industry, the largest in absolute terms, and the three Territorial Authorities that are similar in industry composition.
save_shiny_data.R	This script prepares copies of the data that are used in the deployment of the shinyapp.
deploy_shinyapp.R	This script provides proxy details & permissions for deploying the shinyapp.
create_public_repo_snapshot.R	This script creates a snapshot of the overall project for public release.

Figures created by these scripts are saved in the `dissemination_outputs` folder.

5. Results

5.1 Evaluation of Marginal Totals from Input Data Sources

The script `troubleshoot_incompatible_population_data.R` conducts 6 separate tests for the differences between the marginal totals across the different data sets. Test results are printed to the screen and allow the user to evaluate differences between the marginal totals in absolute and relative terms.

For example, in comparing the publicly released Regional and National GDP sources, all of the differences are less than 1 per cent.

```
tmp_conc <- industries %>%
  select(RGDP_industry, NGDP_industry) %>%
  unique()

comparison <- ngdp_pop %>%
    left_join(tmp_conc) %>%
    group_by(RGDP_industry, Year) %>%
    summarise(NGDP = sum(Freq)) %>%
    left_join(rgdp_pop_pub) %>%
    rename(RGDP = Freq) %>%
    ungroup() %>%
    group_by(RGDP_industry, Year) %>%
    summarise(NGDP = mean(NGDP),
              RGDP = sum(RGDP)) %>%
    mutate(Difference  = RGDP - NGDP,
           DiffPercent = Difference / NGDP * 100) %>%
    ungroup() %>%
    arrange(-abs(DiffPercent))

## Joining by: "NGDP_industry"
## Joining by: c("RGDP_industry", "Year")
```

```
comparison %>% data.frame() %>%
head(5) %>% print()

##                       RGDP_industry Year      NGDP  RGDP Difference
## 1  Accommodation and Food Services 2001  2317.000  2315  -2.000000
## 2  Accommodation and Food Services 2007  3365.117  3363  -2.117312
## 3  Accommodation and Food Services 2011  3857.152  3859   1.848315
## 4 Public Administration and Safety 2001  4506.000  4504  -2.000000
## 5                  Manufacturing 2010 20920.754 20930   9.246099
##   DiffPercent
## 1 -0.08631852
## 2 -0.06291941
## 3  0.04791916
```

```
## 4 -0.04438526
## 5  0.04419582
```

Comparisons among the various sources of Regional GDP showed some differences between the industry and regional margins, which can be attributed in part to the complex industry groupings (e.g. “Forestry, Fishing, Mining, Electricity, Gas, Water and Waster Services”):

```
  tmp_conc <- industries %>%
    select(RGDP_industry, RGDPIndustry_custom) %>%
    unique()

  comparison <- rgdp_pop_custom %>%
    left_join(tmp_conc) %>%
    group_by(RGDP_industry, RGDP_Region, Year) %>%
    summarise(Custom = sum(Freq)) %>%
    left_join(rgdp_pop_pub) %>%
    rename(RGDP = Freq) %>%
    ungroup() %>%
    group_by(RGDP_industry, RGDP_Region, Year) %>%
    summarise(Custom = mean(Custom),
              RGDP = sum(RGDP)) %>%
    mutate(Difference = RGDP - Custom,
           DiffPercent = Difference / Custom * 100) %>%
    ungroup() %>%
    arrange(-abs(DiffPercent))

## Joining by: c("RGDP_industry", "RGDPIndustry_custom")
## Joining by: c("RGDP_industry", "RGDP_Region", "Year")
```

```
 comparison %>% data.frame() %>%
    head(5) %>%
    print() # Not OK

##                                                      RGDP_industry
## 1 Forestry, Fishing, Mining, Electricity, Gas, Water and Waste Services
## 2                                                      Manufacturing
## 3 Forestry, Fishing, Mining, Electricity, Gas, Water and Waste Services
## 4 Forestry, Fishing, Mining, Electricity, Gas, Water and Waste Services
## 5 Forestry, Fishing, Mining, Electricity, Gas, Water and Waste Services
##     RGDP_Region Year Custom RGDP    Difference   DiffPercent
## 1      Gisborne 2006     85   85 -2.842171e-14 -3.343731e-14
## 2   Marlborough 2000    218  218 -5.684342e-14 -2.607496e-14
## 3    Wellington 2006    893  893 -2.273737e-13 -2.546178e-14
## 4 Tasman/Nelson 2012    225  225 -5.684342e-14 -2.526374e-14
## 5 Tasman/Nelson 2013    233  233 -5.684342e-14 -2.439632e-14
```

These tests were also used to validate whether differences could be detected between sources. Such tests can be useful for verifying the content of the downloads from Statistics New Zealand (e.g. if

some industries are selected twice). For example, comparisons between the LEED 18 and LEED 4 tables showed some differences in “Arts and recreation services”, “Mining; electricity, gas, water, and waste services”, and “Rental, hiring, and real estate services”:

```
  tmp_conc <- industries %>%
    select(LEED4Industry, LEED18Industry) %>%
    unique()

# make a national GDP object with RGDP categories
  comparison <- leed4_pop %>%
    left_join(tmp_conc) %>%
    group_by(LEED18Industry, Year) %>%
    summarise(LEED4 = sum(Freq)) %>%
    left_join(leed18_pop) %>%
    rename(LEED18 = Freq) %>%
    ungroup() %>%
    group_by(LEED18Industry, Year) %>%
   summarise(LEED18 = sum(LEED18),       # LEED18 is the one with regions so need to add them up
             LEED4 = mean(LEED4)) %>%
    mutate(Difference = LEED18 - LEED4,
           DiffPercent = Difference / LEED4 * 100) %>%
    ungroup() %>%
    arrange(-abs(DiffPercent))

## Joining by: "LEED4Industry"
## Joining by: c("LEED18Industry", "Year")
```

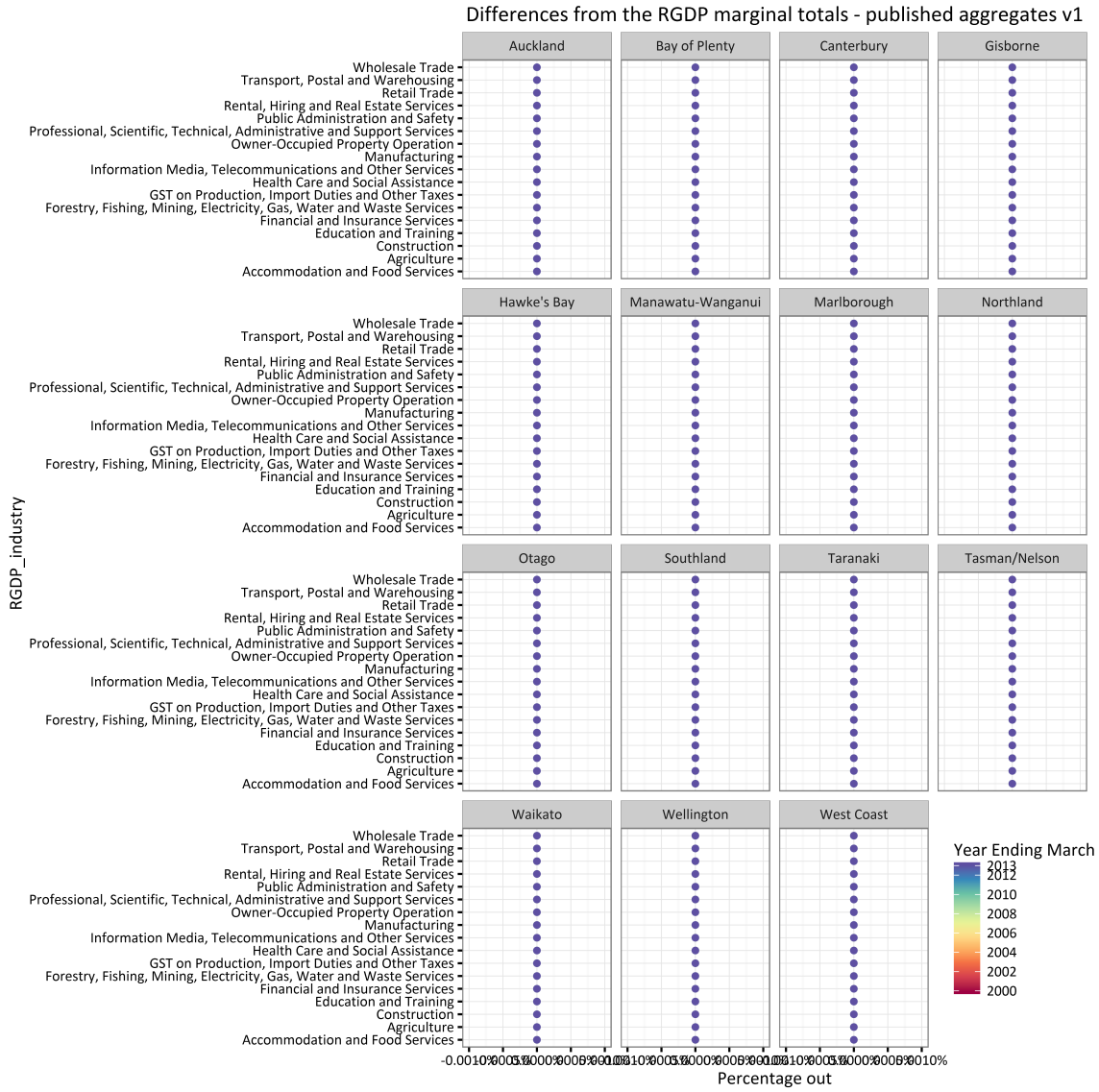
```
comparison %>% data.frame() %>%
   head(5) %>%
   print()

##                                       LEED18Industry Year     LEED18
## 1                         Arts and recreation services 2008 1103765056
## 2 Mining; electricity, gas, water, and waste services 2006  799614340
## 3                         Arts and recreation services 2011 1270155874
## 4                         Arts and recreation services 2004  763363503
## 5            Rental, hiring, and real estate services 2001  633022157
##        LEED4 Difference   DiffPercent
## 1 1103765791  -734.3159 -6.652824e-05
## 2  799613863   477.3375  5.969600e-05
## 3 1270155183   690.3666  5.435293e-05
## 4  763363110   392.7469  5.144955e-05
## 5  633022445  -287.3703 -4.539653e-05
```

5.2 Output Validation

Results from the analysis were validated to look for internal consistencies with the source data (e.g. making direct comparisons from the marginal totals from the LEED tables as well as Regional and

National GDP estimates). Population data that were also incompatible (e.g. matches between published Regional and National GDP figures) were also identified in the screening of results to identify potential sources of error. In general, the output measures for Regional GDP were within 0.05 percent and varied across years (Figure 3). Typically, there were larger deviations from earlier years in the time series (e.g. 2000 to 2006), while later years (i.e. 2012) had little or no variation. There did not appear to have any consistent error across industries.

Figure 3: Validation results of the differences in marginal totals for published regional GDP figures



Comparisons of output with NGDP marginal totals largely varied between 0.5 and these deviations tended to be consistent across industries (Figure 4). For example, TAGDP estimates for 'Agriculture,

Forestry and Fishing Support Services and Hunting', 'Basic Chemical and Chemical Product Manufacturing', and 'Beverage and Tobacco Product Manufacturing' were consistently over-estimated, while 'Fabricated Metal Product Manufacturing', 'Furniture and Other Manufacturing', and 'Machinery and Other Equipment Manufacturing' were consistently under-estimated. As these industries tended to be associated with the same coarser level ANZSIC classification, it appears that the error appears consistent in the finer level ANZSIC groupings.

Figure 4: Validation results of the differences in marginal totals for published national GDP figures

Table 2: Industry totals match those in National GDP (selected industries), 2013

NGDP_industry	Total ($b)
GST on Production, Import Duties and Other Taxes	18.5
Professional, Scientific and Technical Services	16.6
Owner-Occupied Property Operation (National Accounts Only)	14.6
Health Care and Social Assistance	13.1
Property Operators and Real Estate Services	12.7
Wholesale Trade	10.6
Education and Training	10.2
Central Government Administration, Defence and Public Safety	8.1
Finance	6.6
Construction Services	5.9
Electricity and Gas Supply	5.6
Other Store-Based Retailing and Non Store Retailing	5.3
Dairy Cattle Farming	5.0
Postal, Courier Transport Support, and Warehousing Services.	4.5
Administrative and Support Services	4.3
Accommodation and Food Services	4.1
Telecommunications, Internet and Library Services	4.1
Other Services	4.0
Mining	3.7
Arts and Recreation Services	3.1

Table 3: Regional totals match those in Regional GDP, 2013

Region	Total ($b)
Auckland	78.2
Wellington	30.2
Canterbury	27.8
Waikato	18.2
Bay of Plenty	11.4
Otago	9.3
Manawatu-Wanganui	8.7
Taranaki	8.7
Hawke's Bay	6.2
Northland	5.3
Southland	4.9
Nelson	2.4
Marlborough	2.2
West Coast	1.7
Gisborne	1.6
Tasman	1.5

6. Dissemination Outputs

6.1 Comparative Plots for Territorial Authority vs. Industry Comparisons

The MTAGDP project provides code that can be used to re-create outputs used for dissemination of results, including graphics and deployment of a web-based visualisation tool (via shinyapps). One of the principal ways to view the MTAGDP results is to compare industries within Territorial Authorities. The code for creating these graphs are in the `plot_industries_by_TA.R` script, which consists of a loop for all Territorial Authorities (except Chatham Islands). An example of this figure is provided in Figure 5.

Another set of standardised figures can be produced in the `plot_topTAs_by_industry.R` script, which produces plots the Territorial Authority contribution to GDP for a given industry (Figure 6).

6.2 Deployment of the MTAGDP Shinyapp

The deployment of the MTAGDP shinyapp relies on several scripts within the `dissemination_code` folder, including the preparation of the basic commentary used for the "One areas top industries" tab, which identifies the most distinctive and fastest growing industry and reports on GDP measures. The data required for the shiny app are also created in the `save_shiny_data.R` script. Finally, the deployment of the shinyapp defines credentials (as part of MBIE's shinyapps.io account) and deploys the application (named 'mtagdp_test').

The scripts which control the MTAGDP shinyapp functionality are contained in the `shiny` folder. Here the data objects (i.e. `.rda` files) are stored and the `server.R` and `ui.R` which define the graphic displays, panels, and other functionality of the application Figure 7.

Figure 5: Results for a single Territorial Authority

Figure 6: Results for a single detailed industry

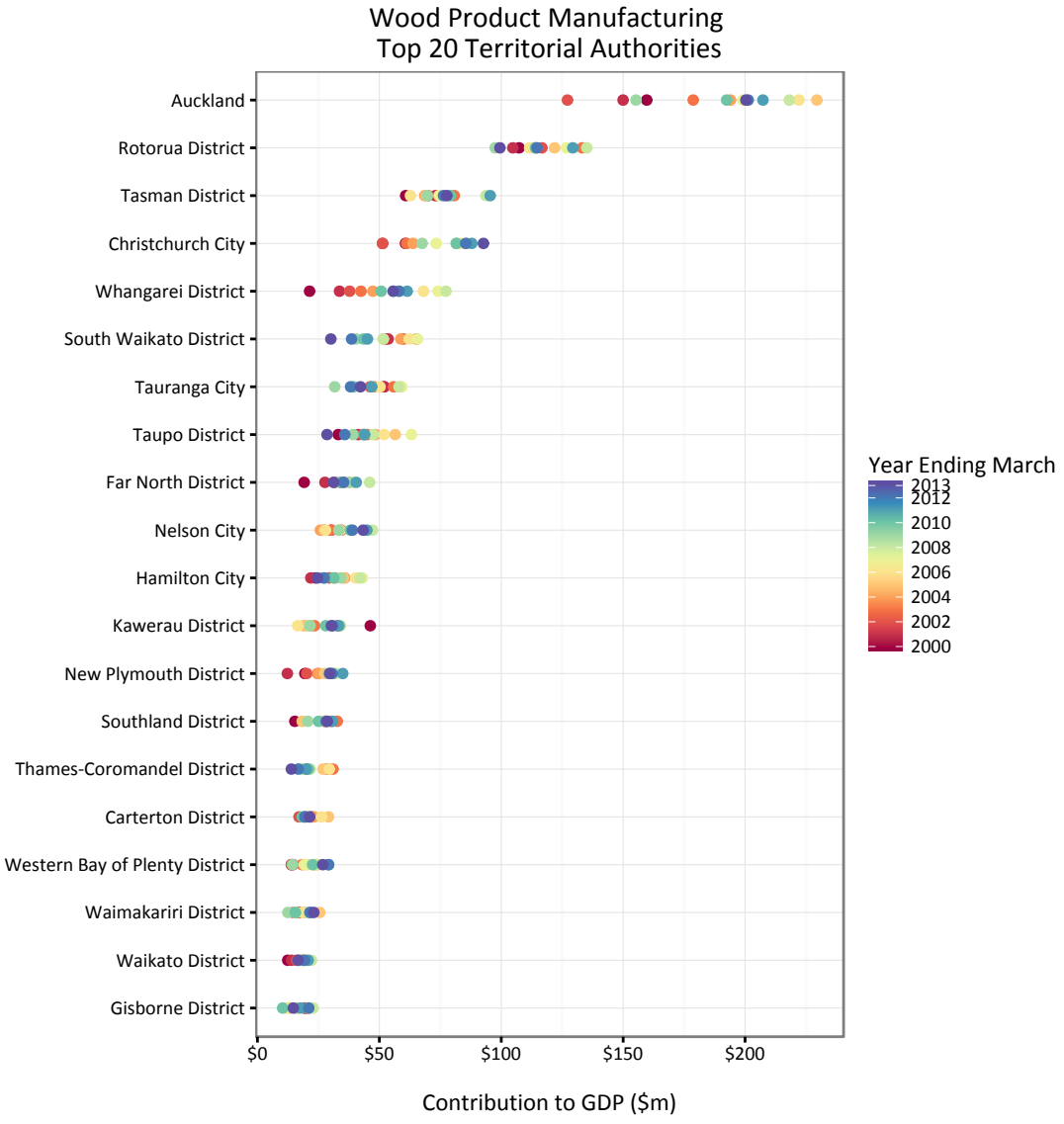
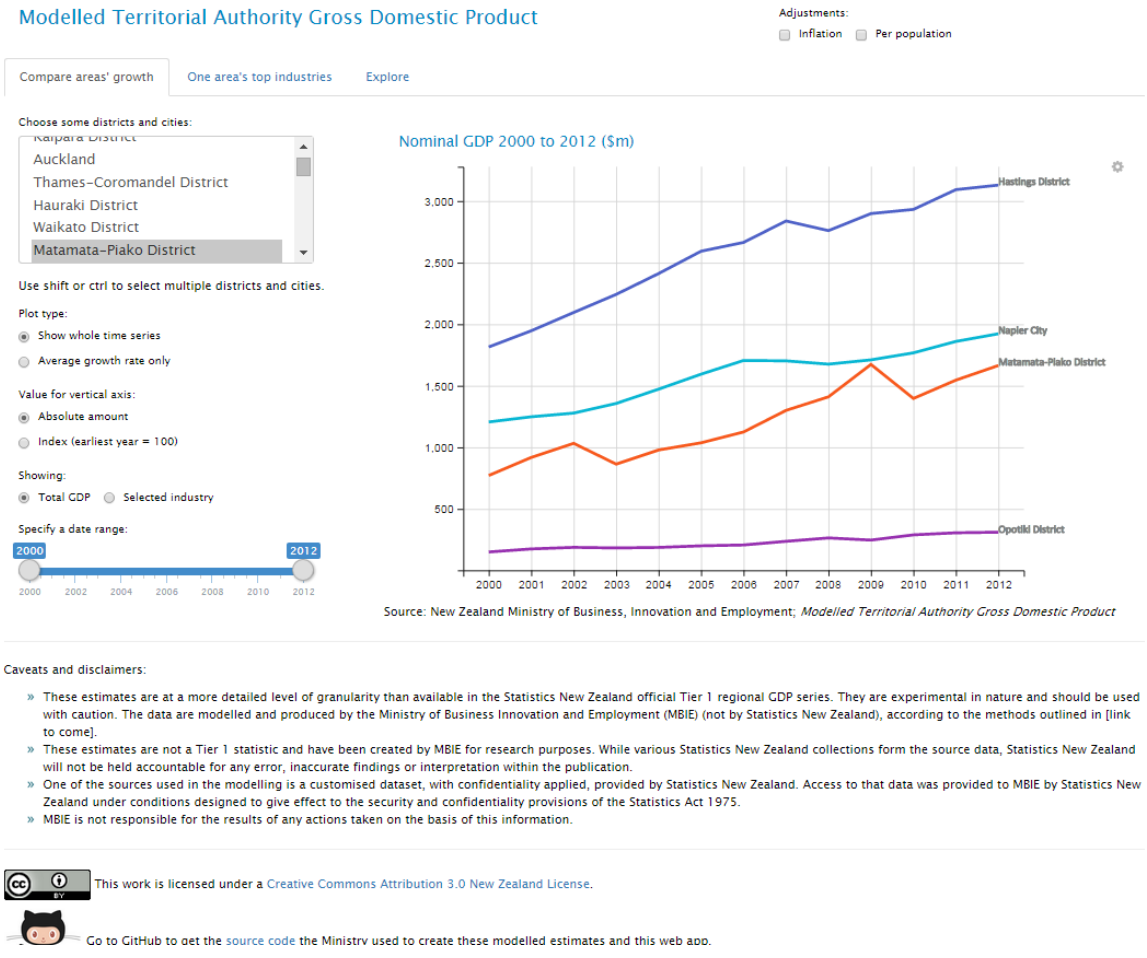


Figure 7: Screenshot of the interactive web tool



7. Assumptions, Caveats and Key Limitations

7.1 Interaction of Industry and Regional Effects

The method has been described in this paper as bottom up, scaling from the BDS jobs numbers up to earnings and then to GDP. The main assumption behind this approach is that ratios of earnings to GDP in different combinations of regions and industries can be determined from the published data in national GDP (i.e. NGDP), regional GDP (i.e. RGDP), and the Linked Employer Employee Database (i.e. LEED). This is a reasonable assumption and reflects the steps of the actual technique.

An alternative conceptualisation of the method is to consider it from the other direction, starting with Regional GDP and seek ways to disaggregate it into more detailed geographical and industry classifications:

- We use the national GDP figures to disaggregate to more detailed industry levels
- We use LEED earnings data to make that disaggregation more precise (i.e. based on the actual split of earnings) to TA level
- When the LEED has reached its limits (does not have the complete tabulation of TA and industry), we use the BDS jobs numbers for the final granularity

In the first instance, when we develop estimates for detailed industry at regional level, all we need to assume is that the relationship from LEED earnings to NGDP industry does not have an interaction effect with Region. This doesn't mean the relationship needs to be the same for each Region and industry - in fact those things changing is the essence of the method - but if one region has a low ratio of GDP to earnings, it has to be low for all the industries. This seems to be a reasonable assumption.

As we go further into the process, these assumptions become more problematic. See Table 4. In identifying TA level estimates we have to assume that the earnings to GDP ratios are the same for each TA within a region. When we develop the final detailed estimates of highly detailed LEED4 industry, we are having to assume that the earnings to GDP ratios for each detailed industry within an NGDP category are equal. For example, the LEED4 categories 'Nursery and floriculture production", "Mushroom and vegetable growing", and "Fruit and tree nut growing" each come under the NGDP category of "Horticulture and Fruit Growing". Using this methodology, we assume the ratio of earnings to GDP in each of these three detailed industries are the same. Although this may seem like a reasonable assumption, further validation on whether this assumption holds (and, for which industries) is required.

As one of the key assumptions for this methodology is that there is no interaction between regional and industry disaggregation of the Regional GDP, future work should also be dedicated to validating this assumption.

7.2 Owner-occupied dwelling operation

Owner-occupied dwelling operation (OOD) - effectively, imputed contribution to GDP from owner-operators involved in selling housing services to themselves - is an important part of the national accounts (depending on how you disaggregate industries, it is usually the largest contributor to GDP). However, it does not feature in either the jobs count in the BDS or the earnings estimates in the LEED. Hence the method described above cannot allocate the Regional OOD figures in RGDP to the Territorial Authorities.

To get around this problem, OOD was distributed to Territorial Authorities in accordance with the relative total earnings in each Territorial Authority. Although this seems like a reasonable assumption, further validation on how 'Owner-occupied dwelling operation' earnings are distributed will be required.

Table 4 shows the increasing uncertainty about the quality of estimates as we get further from the published Tier 1 statistics for Regional GDP and National GDP.

Table 4: Data quality categories for the new TA GDP estimates

Data quality	Data
Published Tier 1 statistics	Region x RGDP Industry National NGDP Industry
Calibrated to a published statistic using LEED earnings - only one step removed	TA x RGDP Industry Region x NGDP Industry
Two steps removed from a published statistic	TA x NGDP Industry Region x LEED4 Industry
Three steps removed - much stronger assumptions about relationship of earnings to GDP	TA x LEED4 Industry

8. Concordances

One of the key elements in the derivation of MTAGDP lies in the concordances of the different levels of industry and geographical area in the published data. This basically allows for linking the different data sources across ANZSIC06 (and other published) industry codes and Regional Council and Territorial Authority boundaries.

8.1 Geographical area

Table 5 shows the relatively straightforward concordance at Regional level between the different data sources.

Table 5: Region concordances (excluding exact matches)

	Region	LEED18Region	RGDP_Region
3	West Coast	Tasman, Nelson, Marlborough, West Coast	West Coast
13	Marlborough	Tasman, Nelson, Marlborough, West Coast	Marlborough
14	Nelson	Tasman, Nelson, Marlborough, West Coast	Tasman/Nelson
16	Tasman	Tasman, Nelson, Marlborough, West Coast	Tasman/Nelson

When matching Territorial Authorities to Regions, a well known problem is the lack of strict hierarchy of classifications. As the original BDS data is at Territorial Authority level but some of the marginal totals to which it is being calibrated are at Regional level, it's necessary to allocate each original observation not only to a Territorial Authority but also to a Region. This needs more granular data than is in the original - for example, Waitomo District is in both Waikato Region and Manawatu-Wanganui Region and hence the number of jobs in the BDS needs to be allocated between the two.

The method employed to meet this problem was to create a new variable, `TA_Region_modified`, that contains 75 sub-district groupings each matched to a single Region. Some Territorial Authorities such as Waitomo are allocated to two different regions. The proportions used are the geographical proportions and were sourced from Wikipedia. These fuzzy concordances are shown in Table 6. Territorial Authorities that are not shown all have one-to-one relations with a Region.

8.2 Industry

The Industry concordance is more straightforward than Regions but still has elements of complexity. No less than five different ANZSIC categorisations are needed - for BDS, LEED4, LEED18, NGDP and RGDP. The first thirty rows of this concordance are shown in Table 7 & 8. The extract in Tables 7 & 8 illustrates several of the complexities dealt with:

Table 6: Territorial Authority concording to multiple Regions

	SNZ_TA	Region	Proportion	TA_Region_modified
13	Waitomo District	Waikato	0.95	Waitomo District 1
14	Waitomo District	Manawatu-Wanganui	0.05	Waitomo District 2
15	Taupo District	Waikato	0.74	Taupo District 1
16	Taupo District	Bay of Plenty	0.14	Taupo District 2
17	Taupo District	Hawke's Bay	0.11	Taupo District 3
18	Taupo District	Manawatu-Wanganui	0.01	Taupo District 4
23	Rotorua District	Bay of Plenty	0.62	Rotorua District 1
24	Rotorua District	Waikato	0.38	Rotorua District 2
32	Stratford District	Taranaki	0.68	Stratford District 1
33	Stratford District	Manawatu-Wanganui	0.32	Stratford District 2
36	Rangitikei District	Manawatu-Wanganui	0.86	Rangitikei District 1
37	Rangitikei District	Hawke's Bay	0.14	Rangitikei District 2
41	Tararua District	Manawatu-Wanganui	0.98	Tararua District 1
42	Tararua District	Wellington	0.02	Tararua District 2
67	Waitaki District	Canterbury	0.60	Waitaki District 1
68	Waitaki District	Otago	0.40	Waitaki District 2

- Aggregations for RGDP such as "Forestry, Fishing, Mining, Electricity, Gas, Water and Waste Services" cross over aggregations made in LEED18 and in NGDP with no strict hierarchy
- Some low level codes available in the BDS are aggregated in LEED4 (eg B060, B070 and B080)

Table 7: Extract from industry concordance - first three columns

	ANZSIC06	ANZSIC06_Description	LEED4Code
1	A011100	Nursery production (under cover)	AA011
2	A011200	Nursery production (outdoors)	AA011
3	A011300	Turf growing	AA011
4	A011400	Floriculture production (under cover)	AA011
5	A011500	Floriculture production (outdoors)	AA011
6	A012100	Mushroom growing	AA012
7	A012200	Vegetable growing (under cover)	AA012
8	A012300	Vegetable growing (outdoors)	AA012
9	A013100	Grape growing	AA013
10	A013200	Kiwifruit growing	AA013
11	A013300	Berry fruit growing	AA013
12	A013400	Apple and pear growing	AA013
13	A013500	Stone fruit growing	AA013
14	A013600	Citrus fruit growing	AA013
15	A013700	Olive growing	AA013
16	A013900	Other fruit and tree nut growing	AA013
17	A014100	Sheep farming (specialised)	AA014
18	A014200	Beef cattle farming (specialised)	AA014
19	A014300	Beef cattle feedlots (specialised)	AA014
20	A014400	Sheep-beef cattle farming	AA014
21	A014500	Grain-sheep or grain-beef cattle farming	AA014
22	A014600	Rice growing	AA014
23	A014900	Other grain growing	AA014
24	A015100	Sugar cane growing	AA015
25	A015200	Cotton growing	AA015
26	A015900	Other crop growing n.e.c.	AA015
27	A016000	Dairy cattle farming	AA016
28	A017100	Poultry farming (meat)	AA017
29	A017200	Poultry farming (eggs)	AA017
30	A018000	Deer farming	AA018

Table 8: Extract from industry concordance - continued

	LEED4Industry	RGDP_industry
1	Nursery and floriculture production	Agriculture
2	Nursery and floriculture production	Agriculture
3	Nursery and floriculture production	Agriculture
4	Nursery and floriculture production	Agriculture
5	Nursery and floriculture production	Agriculture
6	Mushroom and vegetable growing	Agriculture
7	Mushroom and vegetable growing	Agriculture
8	Mushroom and vegetable growing	Agriculture
9	Fruit and tree nut growing	Agriculture
10	Fruit and tree nut growing	Agriculture
11	Fruit and tree nut growing	Agriculture
12	Fruit and tree nut growing	Agriculture
13	Fruit and tree nut growing	Agriculture
14	Fruit and tree nut growing	Agriculture
15	Fruit and tree nut growing	Agriculture
16	Fruit and tree nut growing	Agriculture
17	Grain, sheep, and beef cattle farming	Agriculture
18	Grain, sheep, and beef cattle farming	Agriculture
19	Grain, sheep, and beef cattle farming	Agriculture
20	Grain, sheep, and beef cattle farming	Agriculture
21	Grain, sheep, and beef cattle farming	Agriculture
22	Grain, sheep, and beef cattle farming	Agriculture
23	Grain, sheep, and beef cattle farming	Agriculture
24	Other crop growing	Agriculture
25	Other crop growing	Agriculture
26	Other crop growing	Agriculture
27	Dairy cattle farming	Agriculture
28	Poultry farming	Agriculture
29	Poultry farming	Agriculture
30	Deer farming	Agriculture