DATA471 Group Report

2023-09-02

Authors: Group 15

- Izzy Bremnar ([[]]):
  - Background and data section, and supervisor representative
- Tram Chau ([[]]):
  - EDA section, discord/zoom, kanban and mind maps
- Kim Downing (300 639 199):
  - Ethics, privacy and security section, project manager and editor
- Amish Verma ([[]]):
  - EDA section and git

Background and Data

In our project, we are focusing on a time series dataset which is part of Statistics New Zealand's COVID-19 data portal (1) - specifically we have picked the New Zealand Activity Index (NZAC). The NZAC was created in 2020 by the Treasury (2), but has since been extended to 2004 using historical data. This dataset is a monthly index that provides a summary of the eight constituent datasets which include data on: consumer spending, job vacancies, job-seeker numbers, electricity generation, and traffic data from both public and private institutions such as the Ministry of Social Development, ANZ and SEEK (2).

In this time series index, the eight component datasets are all also numerical. The NZAC is constructed using a weighted average of the components, and aims to show the common movements across the constituent datasets (2). The resultant values are then scaled so that it has the same mean and variance as year-on-year Real Gross Domestic Product growth (as the NZAC is intended as a proxy for the RGDP (2)). This style of activity index became common during the COVID-19 pandemic, when decision-makers in government and business were interested in higher frequency data than has previously been on offer. The NZAC is based on similar indices previously developed by the Chicago and New York Federal Reserves in the United States of America (2). In the United States of America, these measures provide weekly and monthly activity indices for the country (2). The New Zealand Treasury utilises the same weighted average method for the NZAC (2).

The NZAC is important as it is higher frequency than most economic datasets. Measures of GDP and the unemployment rate are released quarterly and can be revised for up to 2 years following release. It is well understood in the economics sector that accurate data can carry a 6-to-18-month lag, especially when considering the whole economy (2). These smaller indexes enable a faster release of data that is more up to date. The NZAC is a particularly useful dataset as it aims to reduce the noise that can be present in the component datasets, if used alone. The NZAC is highly correlated with Real GDP, which means that it could be helpful when making economic forecasts, or if you needed to know how the economy was performing before the quarterly GDP updates are released (2).

The NZAC dataset allows you to tell a clear economic story, especially when you look at how it corresponds to its component datasets and data from different economic sectors e.g., construction or tourism. There are clear features in the data which mark events such as the 2008 Global Financial Crisis and the impacts of the major COVID-19 lockdowns in New Zealand. This data may be of use in addressing questions about how economic activity has changed before, and after COVID. We may be able to look at which components have become more significant, or how the correlation to other datasets has changed.

There are no significant periods of missing date in the main index that we used, however some of the other data that we are using to assess the correlation of the NZAC does have some missing data (see the EDA section for more information). We performed joins on these datasets by using the date column as the join target (after some massaging for one dataset which we will detail in the EDA section).

Ethics, Privacy and Security

Ethics

The dataset is produced by the Treasury and published as part of the Statistics New Zealand COVID-19 data portal. This data is made available under the Crown Copyright, Attribution 4.0 International (CC BY 4.0) license which allows free use, adaptation and distribution of the data as long as the source is attributed, and a link to the license is provided (see earlier in the sentence).

This dataset is derived from constituent datasets that are also publicly available as either full datasets, or as data products. As far as we can tell, none of these datasets have licensing at all - although we believe it is reasonable to expect that Treasury has permission to utilise these datasets and publicly release their derivative product under their own license, which we will respect in using this data.

We also believe that this data was collected with informed consent (as it was mostly either measured values, or through a survey response) and we *assume* that this usage of the data falls within the original bounds of the consent. Here we need to trust that the Treasury has ensured this fact.

The primary dataset has clear ownership, licensing/terms of use, and transparency of where the data came from. As long as we adhere to the terms of the license, it seems difficult to identify any ethical issues in the use of this dataset.

Privacy

This dataset is a series of monthly values which is a combination of multiple datasets. We note that some of the constituent datasets are derived from surveys which may have individual-level, identifying data (although we have not seen these datasets publicly available - these are more likely to be seen as data products). As the final, available data is tightly aggregated, we do not see any potential for de-anonymisation or risk of identifying data being released.

Based on the above, we do not believe there to be any obvious privacy issues with these datasets.

Security

In uploading/updating the dataset on the appropriate website, the organisation needs to ensure that only the necessary people have the ability to access and modify these files. This is an access control mechanism to limit exposure as to who could potentially upload something malicious, or perhaps more likely, reuse a password and let in a blackhat who may choose to do the same.

We also need to be mindful here that the file itself is the correct one, and not some secret or embargoed data, and that it does not accidentally leak metadata (such as who worked on it, commentary from reviewers, exif data in case of images etc).

When the file has been uploaded, the database or server needs to be well defended to deter opportunistic (or targeted) attacks which may aim to change the file (malware, viruses etc - an integrity issue) or deny access to the file (DDOS - an availability issue). For this, we need a properly configured host, which can be a challenge for some (although hopefully not for either the Reserve Bank or Stats New Zealand).

When a user attempts to access/download the file, we must also ensure the integrity of the data from host to client. There are a few ways of doing this, but the most common would be by employing HTTPS in an attempt to thwart man-in-the-middle attacks. Additionally we might provide encryption, or a verification hash. A user might also elect to use a VPN or TOR-like browser, and a firewall and/or anti-virus.

In the specific case of this dataset, StatsNZ offers HTTPS by default, although no verification hash we are aware of. It is difficult to know the specifics of their defenses (which is a good thing). The data is provided in a .csv file, which is a fairly safe filetype given it is pure text data (instead of something like a .pdf or office documents which can utilise macros).

We do not observe any particular security issues related to this dataset.

For our project, the main source of security issues would be in the communication and sharing of files between group members. We have set up a Github repository in order to do this sharing. The repository mitigates many of these potential issues that we might otherwise have in sharing files by email or discord. The repository also features strict access controls in the form of a white-list, and Github itself requires authentication to access (usually in the form of a ssh key or personal access token) which disallows contributions from bad actors (excluding anyone working a Github/Microsoft that is). Although, if one was conniving enough, the information contained in the Github repository may allow someone to do some social engineering, although the attacker would likely have to put in a lot of effort into a (very) low value target.

We do not have any specific counter-measures related to our project outside of good internet hygiene (don't go visiting dodgy sites, or clicking on emails from Nigerian royalty), use a firewall and/or anti-virus, knowing who your group members are, and limiting physical access to your devices.

Exploratory Data Analysis

Data and Preparation

In our analysis, we utilised the primary NZAC data, and four component datasets which included the ANZ Activity Outlook, BusinessNZ's Performance of Manufacturing Index, SEEK Job Ad Index and the SEEK Applications per Ad Index. In preparing the data for analysis, we first needed to:

- Select the relevant columns for each indicator
- Convert the date columns to the 'date' date type
- Ensure all of the dates are the same frequency and date (or massage them to such a state):
    - The ANZ Activity Outlook had date values for the end of each month - we converted these to the first date of the next month for consistency with the rest of the datasets.
    - All datasets have a monthly frequency and all other datasets were dated at the start of the month.

The duration of these datasets are as seen in Table 1 below:

Table 1: Dataset durations

Dataset	Start.Date	End.Date
NZAC	October 2003	June 2023
ANZ Activity Outlook	April 2008	July 2023
BusinessNZ Performance Manufacturing Index	August 2002	June 2023

Dataset	Start.Date	End.Date
SEEK Job Ad Index	January 2002	June 2023
SEEK Applications per Ad Index	Febuary 2008	June 2023

Most of these datasets have coverage between 2008 and 2023, with a few extending past that, back to 2002 or 2003.

Table 2: Summary statistics for the final, merged dataset

Dataset	Variable	Minimum	Median	Mean	Maximum	Missing
All	DATE	2003-10-01	-	-	2023-06-01	0
NZAC	NZAC	-13.605	2.797	2.772	33.801	22
ANZ	ACT_OUTLOOK	-55.10	22.25	18.85	58.50	89
PMI	PMI	26.03	53.35	53.00	64.21	8
PMI	PRODUCTION	19.29	53.92	53.52	68.09	8
PMI	EMPLOYMENT	38.77	50.95	50.55	57.43	8
PMI	NEW_ORDERS	18.25	55.61	54.77	70.48	8
PMI	FINISHED_STOCKS	36.23	51.78	51.77	60.21	8
PMI	DELIVERIES	22.91	53.43	52.69	63.69	8
SEEK Jobs	ADS_SA_INDEX	29.57	113.93	115.89	211.49	1
SEEK Ads	CA_SA_INDEX	28.64	107.7	117.1	284.7	75

In Table 2 above, we show the summary statistics of the resultant dataset after preparation and merging.

First we will state the definition for each of our variables which are shown in Table 2, and will be further used in the EDA below:

[[]]

- the ACT_OUTLOOK…
- PMI
- PRODUCTION
- EMPLOYMENT
- NEW_ORDERS
- FINISHED_STOCKS
- DELIVERIES
- ADS_SA_INDEX
- CA_SA_INDEX

Now looking at content of Table 2 above, our key takeaways are:

- All of the non-date data are numerical
- The different variables have different ranges, with a few variables having negative values, but most are all positive.
- The NZAC in particular, has quite a large outlier at the higher end, with a mean of 2.772 and a maximum value of 33.801.
- None of the variables show a large skew on either side, as the medians and means are fairly close in most respects.
- ACT_OUTLOOK and CA_SA_INDEX had significantly more missing values than compared to other variables. There were no obvious patterns to these missing values as they fluctuated a lot.

Visualisation

Now onto the visualisations. Below we've divided this section into parts:

- Variable correlations
- NZAC vs most correlated plot
- SEEK Ads vs Applications
- NZAC decomposition

We start with the variable correlations - this work was done first (after summary statistics etc) to inform further plots. Our objective is to identify which variables best correlate to our response, NZAC value, and to see whether there may be concerns of multicollinearity to be cognisant of in future modelling stages.

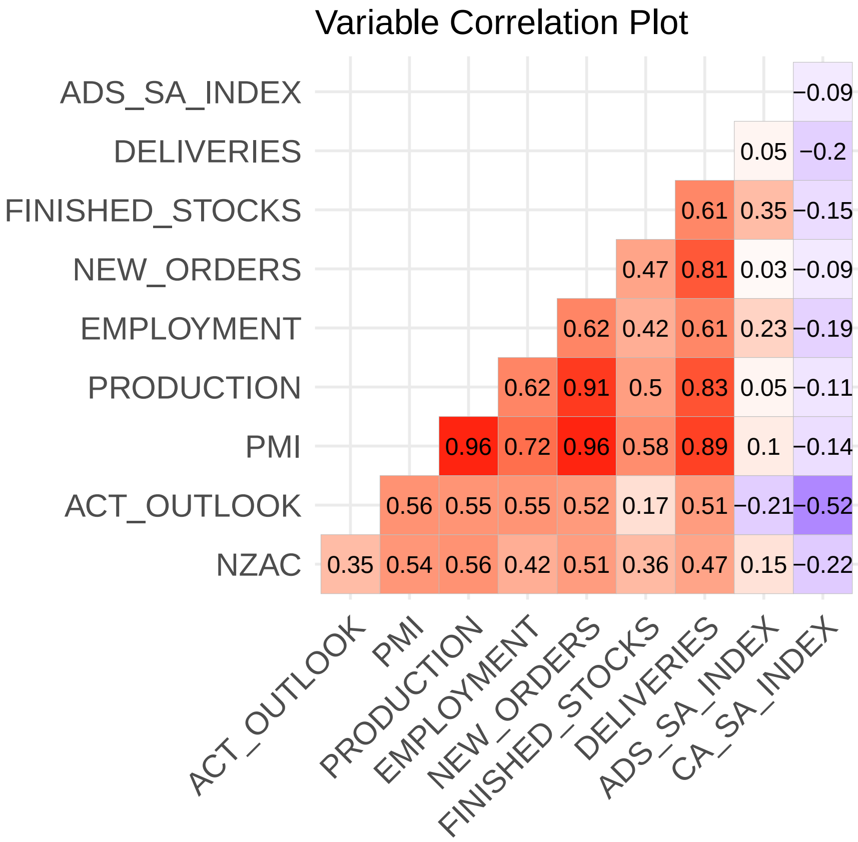

Figure 1: Correlation Plot

First, the reader may want to note that the correlations seen in Figure 1 were made only on cases which had pairwise complete observations between the specific variables as we did have some missing values. We expect that ACT_OUTLOOK and CA_SA_INDEX may not be as accurate as the other variables as they had significantly more missing values. We use the default pearson correlation coefficient here.

The plot itself shows some interesting results:

- The NZAC is moderately correlated to most of the predictors seen here, with correlations between 0.4-0.5 being common.
- CA_SA_INDEX is particularly interesting, as it appears to correlate more to a slower economy (the opposite from the other measures) as being a measure of job applications per job advertisement. This makes sense.

- Consequently, the NZAC has a weak negative correlation with CA_SA_INDEX at -0.22.
- PRODUCTION best correlates to the NZAC at 0.56, while ADS_SA_INDEX correlated the least at 0.15.
- Additionally, we observe the potential for multicollinearity between PMI, PRODUCTION, NEW_ORDERS and potentially DELIVERIES. This is unsurprising given that these measures were all from the same dataset, and all aim to measure similar things, i.e., all new orders to manufacturers will need to be produced, and subsequently delivered.

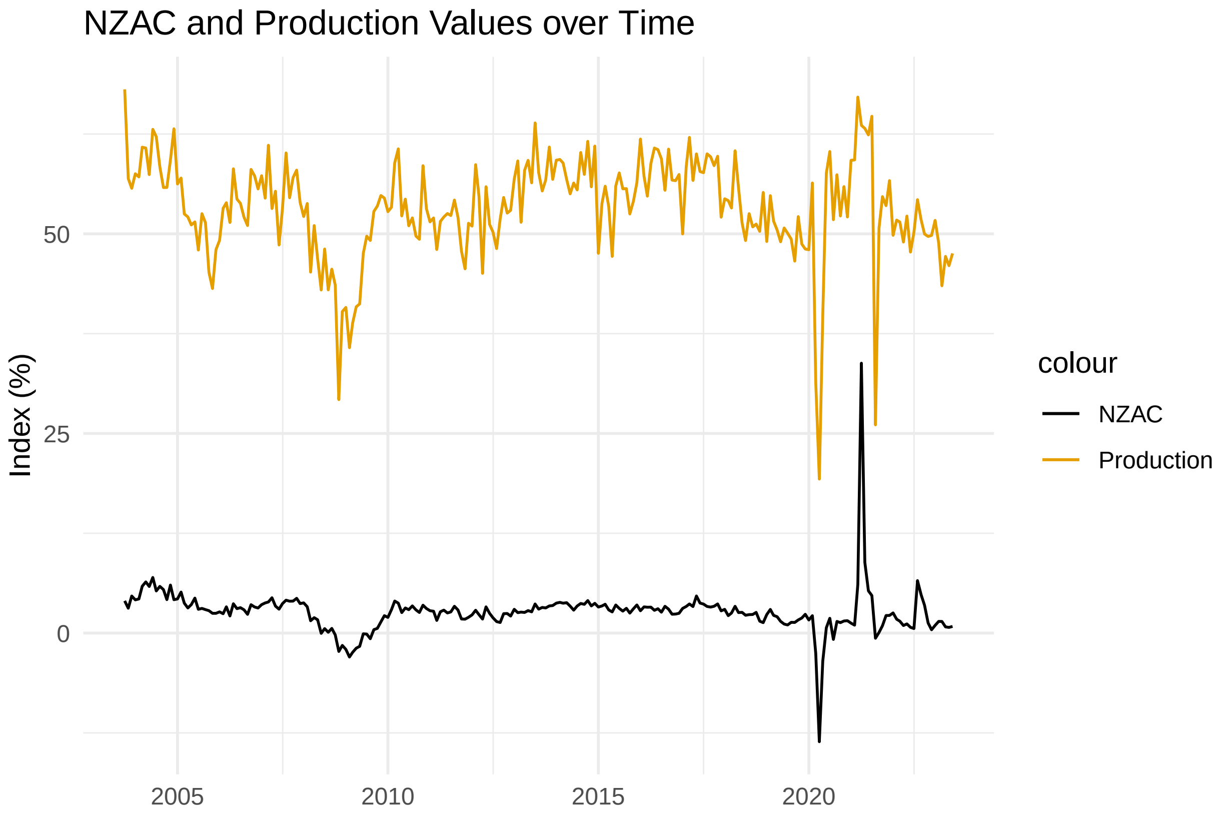

Figure 2: NZAC and Production Values

The NZAC and PRODUCTION value plot seen in Figure 2 shows the NZAC in blank, and the PRODUCTION variable in yellow. NZAC is fairly table over the long-term, with some obvious deviations at key points in time - in the 2007/2008 Global Financial Crisis, and in 2020/2021 due to COVID-19. Interestingly, the shapes of the peaks are very different. In the GFC, there was a longer, but shallower decrease in the NZAC, while COVID produced a short, sharp negative peak - followed, about a year later, by a very large, but still short and sharp, upward deviation. Otherwise, this dataset is largely stable over time.

For the PRODUCTION variable, there are clear signs of correlation visible - similar peaks can be seen during the small early peaks of 2004, the decreases in the GFC and the significant drop due to onset of COVID lockdowns. Interestingly, there is a slight peak coinciding with the large increase noted above somewhere into COVID, but not by as much. This may make sense as there may only be so much capacity that can be filled in a short time, and there may need to be additional infrastructure and staffing required to ramp up productivity much more than that, which cannot be done in that short of a time.

The SEEK job ads volumes and job applications per ad graph seen in Figure 3 tells a similar story. For the number of listed jobs, we see similar dips in the GFC and in early COVID, with a quick and resounding

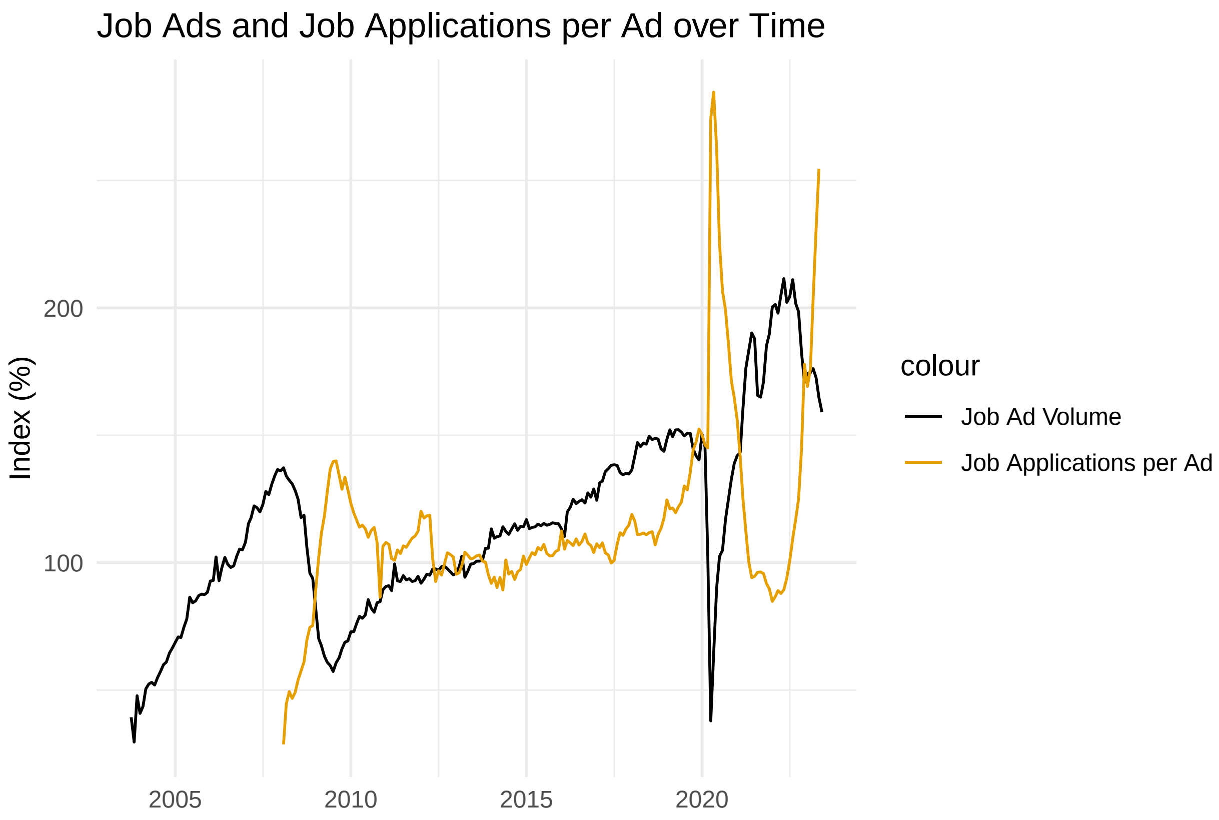

Figure 3: SEEK Jobs data: Job Ads and Job Applications

rebound about a year into COVID. A key difference in this plot is that there is a general upwards trend since 2004.

The job applications per ad tell the same story, although perhaps from the other side of the coin. It makes intuitive sense that when there are less jobs, there will be more people competing for them, and when there are more jobs, there will be less demand. These variables together tell this story, and gives us reassurance and confidence in these points, and the effect that COVID had on job listing and applications.

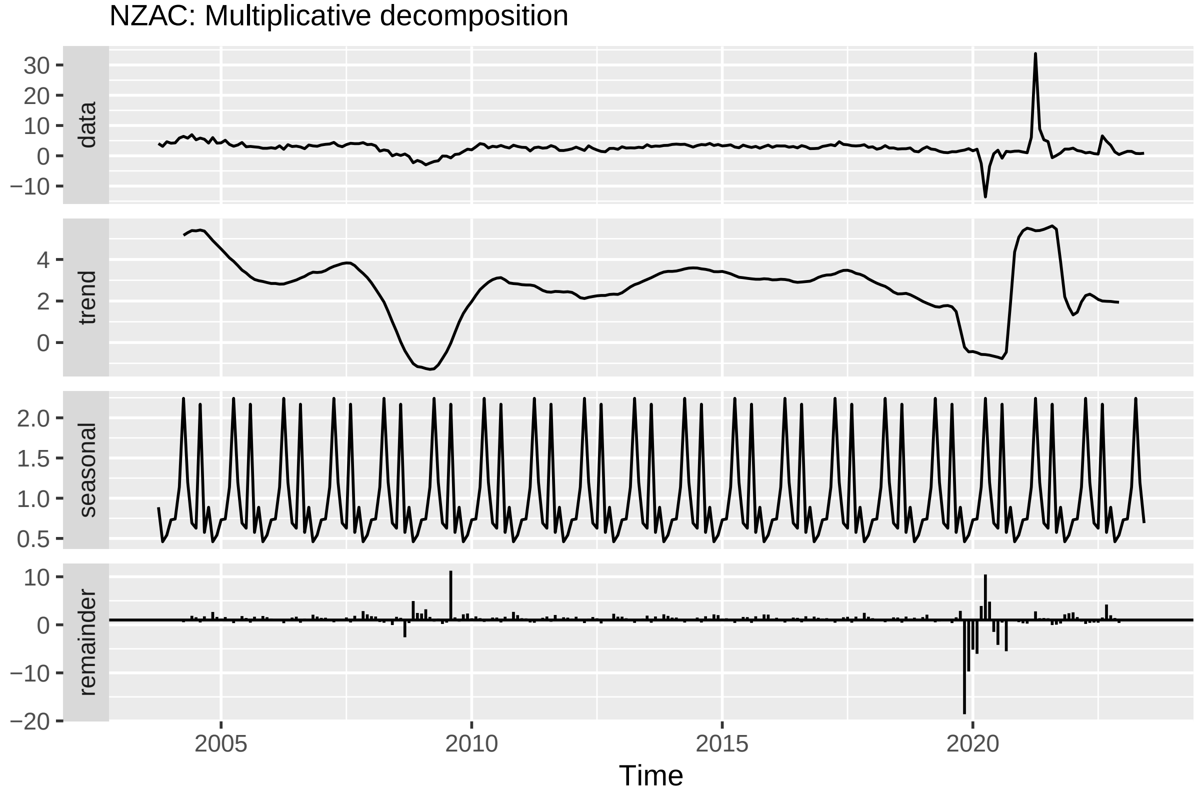

Figure 4: Temporal decomposition plot

This decomposition graph above in Figure 4 shows the effects of different timings on our data. First, the data itself is plotted above at the top, follow by the underlying trend, seasonality and whatever remains at the bottom (unexplained error). As noted previously, we see the effects of the GFC and COVID clearly. Additionally we see a long-term trend of about 2 units, and a roughly yearly seasonal effect of between 0.5-2 units, which appears to have two large peaks, with a smaller third. The real meat of this graph is in the remainders - there are large unexplained peaks, both positive and negative, which is understandable as this decomposition apparently could not predict the GFC, nor COVID. The data shows that the NZAC index gradually decreased during 2008, bottoming out at around -3%, while during the COVID-era, the index sharply dropped to -10%. There was a significant post-COVID recovery in 2021 which peaked at over 30% (as compared with previous years drop of -10%).

References

1. COVID-19 data portal [Internet]. Statistics New Zealand; 2020 [cited 2023 Sep 2]. Available from: https://www.stats.govt.nz/experimental/covid-19-data-portal

2. New zealand activity index (NZAC): Technical note [Internet]. The Treasury; 2020 [cited 2023 Sep 2]. Available from: https://www.treasury.govt.nz/publications/nzac/nzac-technical-note