Catalogue no. 11-633-X — No. 027
ISSN 2371-3429
ISBN 978-0-660-35713-3

Analytical Studies: Methods and References

Experimental Economic Activity Indexes for Canadian Provinces and Territories: Experimental Measures Based on Combinations of Monthly Time Series

by Nada Habli, Ryan Macdonald and Jesse Tweedle

Release date: August 19, 2020

Statistics Canada Statistique Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

Email at STATCAN.infostats-infostats.STATCAN@canada.ca

Telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following numbers:

• Statistical Information Service	1-800-263-1136
• National telecommunications device for the hearing impaired	1-800-363-7629
• Fax line	1-514-283-9350

Depository Services Program

• Inquiries line	1-800-635-7943
• Fax line	1-800-565-7757

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "Contact us" > "Standards of service to the public."

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Published by authority of the Minister responsible for Statistics Canada

An HTML version is also available.

Cette publication est aussi disponible en français.

Experimental Economic Activity Indexes for Canadian Provinces and Territories: Experimental Measures Based on Combinations of Monthly Time Series

by

Nada Habli, Ryan Macdonald and **Jesse Tweedle**
Economic Analysis Division,
Statistics Canada

11-633-X No. 027
2020003
ISSN 2371-3429
ISBN 978-0-660-35713-3

August 2020

Analytical Studies: Methods and References

Papers in this series provide background discussions of the methods used to develop data for economic, health, and social analytical studies at Statistics Canada. They are intended to provide readers with information on the statistical methods, standards and definitions used to develop databases for research purposes. All papers in this series have undergone peer and institutional review to ensure that they conform to Statistics Canada's mandate and adhere to generally accepted standards of good professional practice.

The papers can be downloaded free at www.statcan.gc.ca**.**

Acknowledgements

The authors would like to express their thanks to Etienne Saint-Pierre, Yvan Clermont and Danny Leung and Brenda Bugge for their support and guidance during the course of the project. They would also like to thank Steve Matthews, Fred Picard and Michel Ferland for their comments and input on modelling. They also wish to thank Philip Smith and Karen Wilson for their advice, suggestions and support.

Table of contents

Abstract

This paper explores methods for creating a monthly indicator of economic activity for the provinces and territories. It begins by constructing a dataset for the provinces and territories composed of monthly series about the labour force, including wages and employment; international trade; output measures such as manufacturing sales or electricity production; and, prices (consumer, housing and electricity). Where necessary, the series are seasonally adjusted, linked and deflated to create continuous time series from January 2002 to April 2020. Variable reduction methods are then applied to the monthly provincial and territorial dataset to create the experimental monthly provincial and territorial economic indicator indexes. Three methods are examined: Principal components analysis (PCA), Least absolute shrinkage and selection operator (LASSO) and a simple index comprised of three pre-determined series (total employment, total exports and retail sales). A weighted average of the simple index and the PCA index is also constructed. In general, the indexes produced track provincial economic activity reasonably well, following cyclical movements of provincial and territorial economies. However, the set-up for the models is not ideal as annual data are used to produce model parameters, and this leads to uncertainty about model performance. As a result, multiple indexes are reported. A quality assessment provides an indication of the strengths and limitations of the indexes with respect to different uses.

1 Introduction

Timely measures of economic activity are critical for understanding how economies perform, and for informing policy responses to macroeconomic fluctuations. The onset of the pandemic due to the emergence of the SARS-Cov-2 virus emphasized this, as well as the need for geography-specific measures. Presently, Canada has a robust system for producing up-to-date measures of activity, such as real gross domestic product (GDP), at the national level. For provincial and territorial economies, monthly information on labour markets or particular activities such as manufacturing or international trade are available, but a monthly measure of aggregate economic activity is not available.

Under normal circumstances, producing a new set of aggregate economic indicators for the provinces and territories would require creating exploratory measures, possibly launching new surveys or expansions of existing statistical collection activities, as well as creating the infrastructure necessary to produce and disseminate the indicators on an ongoing basis. These changes require time to implement. In the current context, where the SARS-Cov-2 pandemic is accentuating needs for monthly regional economic indicators, the time necessary to constitute a new statistical program to meet current requirements means that this approach is unfeasible.

A more timely approach is to adopt a statistical-model based strategy to quickly create exploratory measures of provincial economic activity. While this approach introduces a new measure of economic activity in a timely fashion, the trade-off is that the methods employed are not used in their ideal situations, that inputs for models use currently available series without an ability to tailor their uses to the creation of an indicator, and that the models are somewhat a-theoretic. That is, the models look for correlations in data rather than employing economic theory to help guide their construction. And, lastly, the models employed here typically have a different set of inputs for each province or territory. As a result, consistency in model structures cannot be maintained across provinces and territories, and this may affect inter-jurisdictional comparisons.[1]

To estimate the activity indexes, a twofold strategy is applied. First, a balanced panel data set of monthly provincial and territorial time series is constructed from publicly available Common Output Data Repository (CODR) tables. This data set spans January 2002 to the latest available data points (currently April 2020). Second, three methods for transforming the monthly series into an indicator of provincial and territorial economic activity are applied to the data set: a simple model, Principal components analysis (PCA) and Least absolute shrinkage and selection estimation (LASSO). A fourth index that combines the index from the simple model and the index based on PCA is also produced. The results from these approaches form the exploratory measures of provincial aggregate economic activity presented here.

The experimental indexes are based on the use of statistical models with an input data set that contains a series of approximations and assumptions. Notable among the assumptions for the input data set are: the use of national deflators to produce real provincial series when provincial

[1] It is unclear to what extent the differences in the number of inputs and diversity of inputs may affect comparisons. Larger provinces have a richer set of data to draw upon when modelling their economies. As well, the surveys that collect provincial and territorial data typically have more observations and/or produce data that has less variability relative to trend movements for larger jurisdictions. Smaller economies are also less complex, and may require a smaller number of series to summarize their activities.
This situation is normal when making comparisons across economies. For example, larger economies in the Organization for Economic Cooperation and Development (OECD) such as the United States, Germany, the United Kingdom or Canada can have bigger statistical systems than smaller OECD economies. This does not make the data from these economies incomparable. Rather, studies are undertaken with an understanding that the data are fit for making comparisons as they reflect what is happening within the jurisdictions they are intended to report on even if the quantity of statistical information available differs. The same notion should be applied to the experimental economic activity indexes.

deflators do not exist, the assumption that the growth rates for all series are covariance stationary, and the assumption that winsorizing data at the 5^{th} and 95^{th} percentiles is appropriate.

For the models, annual real GDP growth across the provinces and territories is used as a measure of aggregate provincial activity against which the derived index measures of economic activity are compared. This produces a situation where a small number of annual observations are being used to estimate model parameters that are used to infer monthly fluctuations. The small number of observations affects the ability of models to provide sufficient inference, and the use of annual data may mask important differences in the monthly timing of changes in prices, outputs and employment.

To allow for the possibility that using annual data may affect model performance and inference, the simple model assumes a set of variables is appropriate and uses OLS to determine their relative contributions. For PCA, a maximized adjusted R-squared from OLS regressions is used to select which of the first ten principal components should be included. In both of these situations the OLS regressions include variables that are statistically insignificant. For LASSO, statistical significance of potential input variables determines the final model. However, for LASSO, for some economies, no input variables are selected. In these cases, elastic net or a general-to-specific modelling strategy is employed instead.

Since the input data set contains a number of assumptions and approximations, and since the statistical models are used in an imperfect setting, the experimental indexes that are created should be viewed as approximations to aggregate economic activity rather than as exact measures. The activity indexes tend to present considerable monthly volatility, and when compared with the real GDP estimates produced by the Government of Quebec or the Government of Ontario, they can exhibit greater cyclical volatility.

Across the measures, the simple models and the LASSO models tend to rely more on employment series as inputs. The simple index has the strength that it is straightforward to understand, and is the most comparable across provinces and territories since it holds the variables used in the activity index constant regardless of the jurisdiction. The PCA indexes appear to capture more fluctuations related to other aspects of overall activity (e.g. sales or exports), but they also have greater variability. The weighted combination of the simple, researcher-defined indexes and the PCA indexes has a better correlation with real GDP growth than the constituent indexes.

Currently, the weighted indexes or the LASSO based indexes appear to offer the best trade-off between signals present in the data and variability of monthly series. This assessment is based on how well the models appear to conform with the set-up used for estimation as well as on the behavior of the indexes. In some cases, such as the PCA based index for Newfoundland and Labrador, an anomaly is present that calls the veracity of the index into question. When these types of situations occur, the data are deemed to be unfit for use at the present time, and the estimates for that index are not provided. The assessments are ongoing and may lead to changes in recommended uses or indexes as further development of the input data set, model refinements or alternative model strategies are explored.

The remainder of this paper is structured as follows. Section 2 discusses the creation of the input data set as well as the assumptions employed to filter and transform the data prior to modelling. Section 3 describes the models employed, the assumptions embedded in the models, as well as their strengths, their limitations and their application for creating monthly indexes. Section 4 provides analysis of the model performance and illustrates the resulting indexes. Section 5 concludes.

2 Input data set

The input data set is comprised of province- and territory-specific measures for economic activity and Canada-level deflators, except in instances where province- and territory-specific deflators are available. The monthly input series are comprised of monthly surveys for labour, outputs and prices (Table 1). In some cases, active tables do not contain continuous information from January, 2002 to the present. In these cases historical tables are used to backcast active tables.

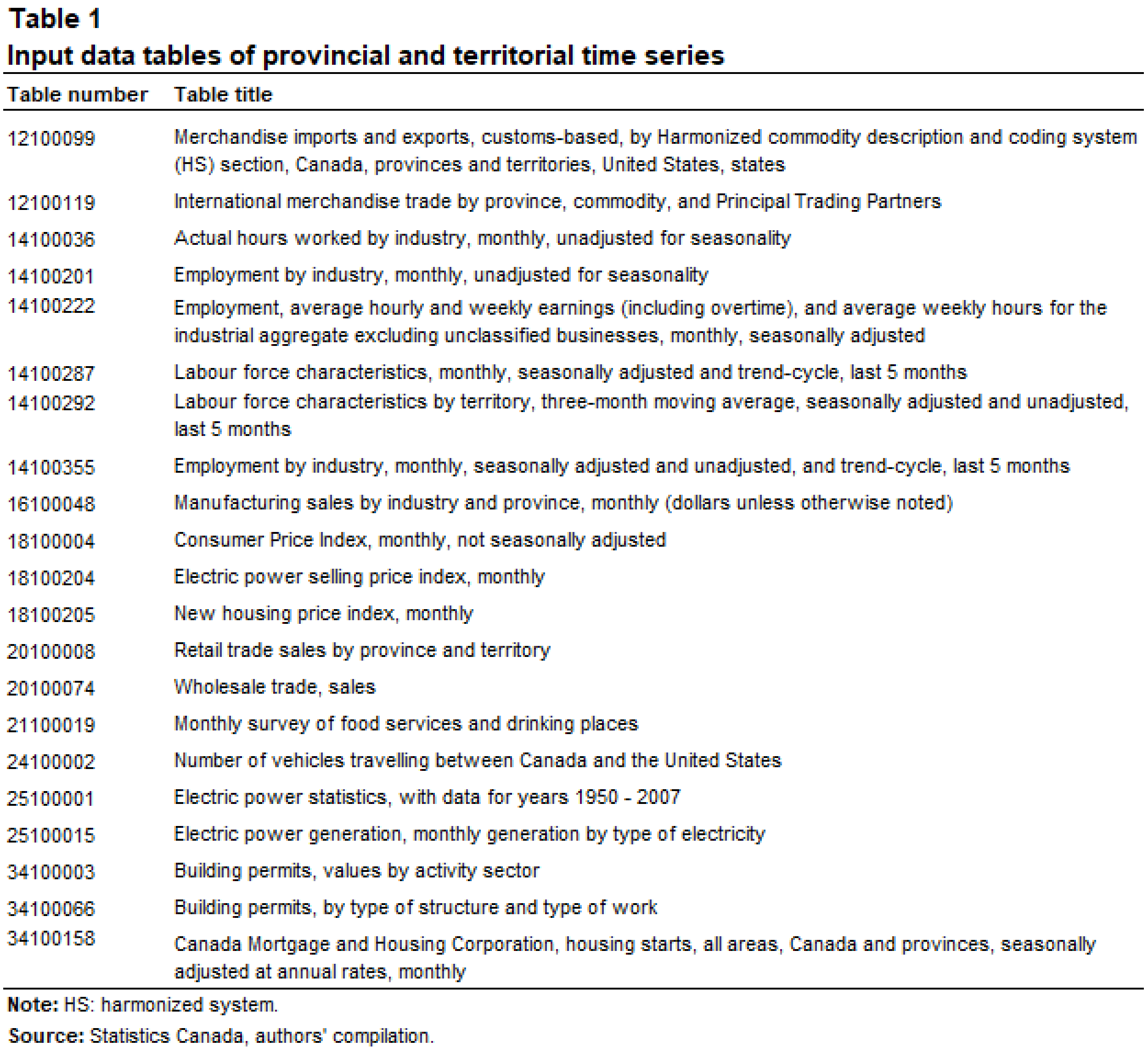


Table 1
Input data tables of provincial and territorial time series

Table number	Table title
12100099	Merchandise imports and exports, customs-based, by Harmonized commodity description and coding system (HS) section, Canada, provinces and territories, United States, states
12100119	International merchandise trade by province, commodity, and Principal Trading Partners
14100036	Actual hours worked by industry, monthly, unadjusted for seasonality
14100201	Employment by industry, monthly, unadjusted for seasonality
14100222	Employment, average hourly and weekly earnings (including overtime), and average weekly hours for the industrial aggregate excluding unclassified businesses, monthly, seasonally adjusted
14100287	Labour force characteristics, monthly, seasonally adjusted and trend-cycle, last 5 months
14100292	Labour force characteristics by territory, three-month moving average, seasonally adjusted and unadjusted, last 5 months
14100355	Employment by industry, monthly, seasonally adjusted and unadjusted, and trend-cycle, last 5 months
16100048	Manufacturing sales by industry and province, monthly (dollars unless otherwise noted)
18100004	Consumer Price Index, monthly, not seasonally adjusted
18100204	Electric power selling price index, monthly
18100205	New housing price index, monthly
20100008	Retail trade sales by province and territory
20100074	Wholesale trade, sales
21100019	Monthly survey of food services and drinking places
24100002	Number of vehicles travelling between Canada and the United States
25100001	Electric power statistics, with data for years 1950 - 2007
25100015	Electric power generation, monthly generation by type of electricity
34100003	Building permits, values by activity sector
34100066	Building permits, by type of structure and type of work
34100158	Canada Mortgage and Housing Corporation, housing starts, all areas, Canada and provinces, seasonally adjusted at annual rates, monthly

Note: HS: harmonized system.

Source: Statistics Canada, authors' compilation.

Deflators are primarily collected from Canada-level survey programs for measures of economic activity (Table 2). Statistics Canada does not currently produce province- and territory-specific deflators for current dollar measures of international trade, manufacturing sales, wholesale sales, retail sales or the Monthly Survey of Food Services. To collect deflators, Canada-level price indexes are taken from surveys when they are available. In the case of manufacturing, the implicit price index is derived as the ratio of the nominal value to the real value. Deflators exist for nominal series back to January 2002, except for manufacturing. For manufacturing, the Index Produce Price Index (IPPI) by industry, IPPI by product group, and the ratio of nominal to real monthly GDP are used as projectors for manufacturing deflators.

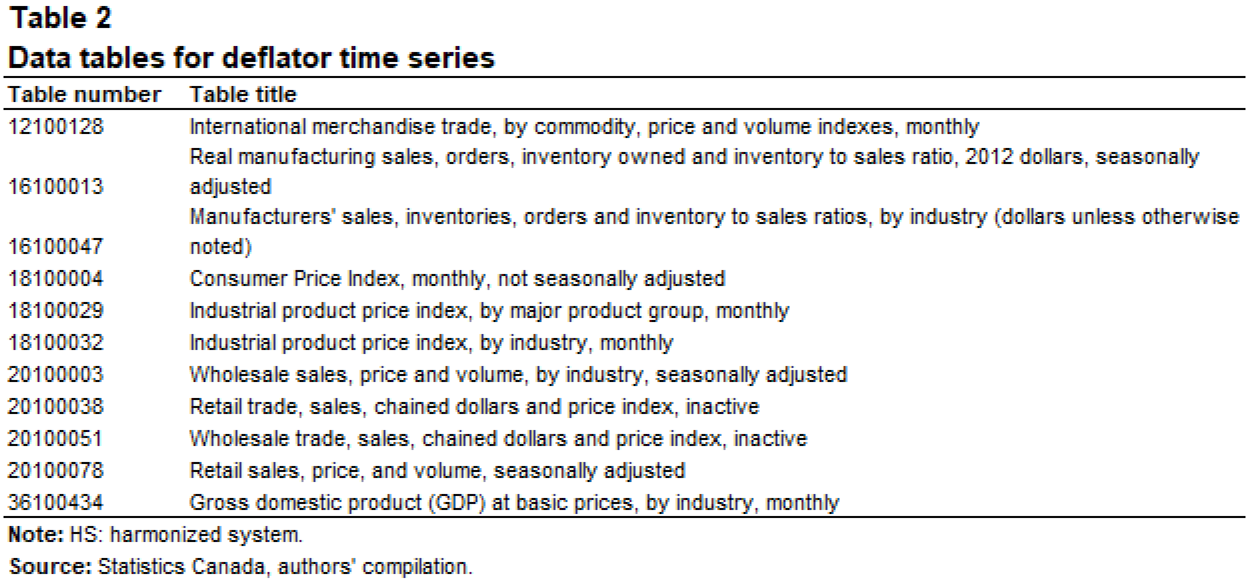


Table 2
Data tables for deflator time series

Table number	Table title
12100128	International merchandise trade, by commodity, price and volume indexes, monthly
16100013	Real manufacturing sales, orders, inventory owned and inventory to sales ratio, 2012 dollars, seasonally adjusted
16100047	Manufacturers' sales, inventories, orders and inventory to sales ratios, by industry (dollars unless otherwise noted)
18100004	Consumer Price Index, monthly, not seasonally adjusted
18100029	Industrial product price index, by major product group, monthly
18100032	Industrial product price index, by industry, monthly
20100003	Wholesale sales, price and volume, by industry, seasonally adjusted
20100038	Retail trade, sales, chained dollars and price index, inactive
20100051	Wholesale trade, sales, chained dollars and price index, inactive
20100078	Retail sales, price, and volume, seasonally adjusted
36100434	Gross domestic product (GDP) at basic prices, by industry, monthly

Note: HS: harmonized system.

Source: Statistics Canada, authors' compilation.

To combine the provincial data and deflator data to create the input data set, there are 4 steps:

1. Assemble data. The CODR tables are filtered to select the desired data. Only variables with continuous data are selected. Series subject to suppression are excluded, however series with 0 values are included. Series subject to suppression are typically smaller value series, meaning they contain less information for aggregate economic fluctuations. Although methods exist for interpolating these data points, should the suppression occur in the latest month, a forecast would be required to infill the suppressed data point. Given that the index is being constructed to provide information about the largest shock to affect the Canadian economy since World War 2, it is considered inadvisable to include series with forecasted values.

2. Seasonally adjust data series. Not all series are provided on a seasonally adjusted basis. There are a total of 966 series seasonally adjusted for use in the indicator. Given the high number of series, the auto options of ARIMA-SEATS algorithm from the R package Seasonal is employed to remove seasonality. Seasonal parameters are determined based on the available monthly time series up to December 2019. In the event the time series do not span the whole period, such as discontinued series, the data up to the most recent period are used to determine the seasonal adjustment options.

   In order to ensure that the seasonally adjusted series are of good quality, they were validated by a range of quality measures. This includes checking for the presence of seasonality, the amount of stable seasonality present relative to the amount of moving seasonality (M7), the absence of seasonal effects in the irregular component, the smoothness of the seasonally adjusted series against its raw form and controlling the number of outliers auto detected by ARIMA-SEATS to a maximum of five. Series that were found to be non-seasonal (125 series) are kept in their raw form. Seasonally adjusted series with poor quality (183 series) are filtered out of the data set prior to estimation.

3. -Link data. After seasonal adjustment, data are linked as necessary. When overlapping periods exist, links are made by treating data as indexes and chaining backwards over historical periods. If overlapping periods do not exist, level values are joined "as is".

4. Apply deflators. Where necessary, deflators are applied to current dollar series or to price series to produce relative price variables.

The full data set prior to filtering contains province- and territory-specific seasonally adjusted and not-seasonally adjusted series, nominal series and deflated series. For modelling provincial and territorial economic activity, series in natural units (e.g. employees, hours worked), deflated series, rates (e.g. the unemployment rate), and relative prices are selected.

3 Estimation

The objective is to estimate monthly time series for aggregate economic activity in the provinces and territories as a function of available monthly provincial and territorial economic time series:

$$y_{monthly,t} = f\left(monthly/time/series\right)$$

The most commonly cited measure of aggregate economic activity is the real GDP measure described in the 2008 System of National Accounts (United Nations 2010).[2] For the Canada-wide level, the function for transforming input series into monthly real GDP is based on industry-specific methodologies and benchmarks that have been developed and enriched over time. The methodologies use a number of data sources to estimate changes in gross output that serve as proxies for real GDP. The proxies are combined with annual real GDP benchmarks to produce the monthly real GDP series. In many cases, a direct measure of gross output is available. However, in some industries, direct measures of output are not available and estimates are constructed using alternative data sources, such as employment (Statistics Canada 2020b).[3] This methodology forms the function $f\left(.\right)$ into which the monthly series are placed to produce monthly real GDP for Canada.

The challenge for measuring provincial and territorial aggregate economic activity is that the function $f\left(.\right)$ for the provinces and territories is unknown, and that the desired series, $y_{monthly,t}$ is also unknown.

Since the goal is to estimate $y_{monthly,t}$, if a close substitute existed, it could be used as an instrument for the true $y_{monthly,t}$. The function $f\left(.\right)$ for combining monthly information to produce an aggregate measure of economic activity could then be approximated. While a close monthly substitute for $y_{monthly,t}$ does not exist, the Canadian System of Macroeconomic Accounts does produce a measure of provincial and territorial GDP at an annual frequency. The approach followed here, therefore, assumes that the annual data can be used as an instrument to help

[2] The Canadian System of Macroeconomic Accounts produces annual, quarterly and monthly measures of GDP for Canada. The measures are integrated and monthly measures of real GDP are derived based on a methodology that uses pseudo-output indexes at a monthly frequency. These indexes are weighted together using index number formulas to form the aggregate monthly real GDP series. The indexes are benchmarked to the annual supply and use tables each year, and this corrects for differences that may arise from the use of pseudo-output indexes in place of the double-deflated value added series produced in the supply and use tables.

[3] Detailed methodologies for individual industries as well as discussions about pseudo-indexes and their use can be found at Canada (2020a).

inform the structure of $f(.)$ when the monthly growth rates from the available input series are averaged within a calendar year and used to estimate parameter values. The parameters of $f(.)$ and the variance characteristics of the monthly input series are then adjusted to account for the difference in periodicity. Monthly indexes of aggregate economic activity are then constructed from the estimated monthly values of $y_{monthly,t}$.

Using a lower frequency variable as the instrument for monthly economic activity lowers the number of degrees for freedom and introduces issues related to the timing of monthly versus annual fluctuations. These issues will have consequences for the ability of the models to produce a monthly estimate of aggregate economic activity. The small degrees of freedom and the covariance around the 2008 recession will tend to produce statistically insignificant parameters for $f(.)$, if the regressors are not importantly affected by business cycle fluctuations. Moreover, models may tend toward selecting a smaller number of inputs at an annual frequency than is necessary for explaining monthly variation as important monthly fluctuations may be masked through aggregation to a lower frequency. Additionally, changes related to prices, sales/output and employment will occur contemporaneously at an annual frequency. However, at a monthly frequency these fluctuations may not align.

Given that a conversion from annual to monthly frequency is necessary to generate the desired estimated values, the modelling strategy includes some approaches that err on capturing more variation in the data rather than focusing solely on model parsimony. This does not mean that models are produced in an ad-hoc fashion. Rather, selection criteria, such as maximizing an adjusted R-squared, are employed alongside more traditional general-to-specific-type modelling strategies.

In summary, because the functional form for transforming monthly data into an aggregate measures of economic activity is unknown, and because the actual values for the series $y_{monthly,t}$ are also unknown, the best that is possible is to approximate the true $y_{monthly,t}$. This means that the series $\hat{y}_{monthly,t}$ will have the flavor for what a real GDP series could look like, but it will not be a true measure of monthly real monthly GDP. Instead, it will be an estimate for an economic activity index which corresponds to macroeconomic conditions in the provinces and territories.

3.1 Estimation strategy

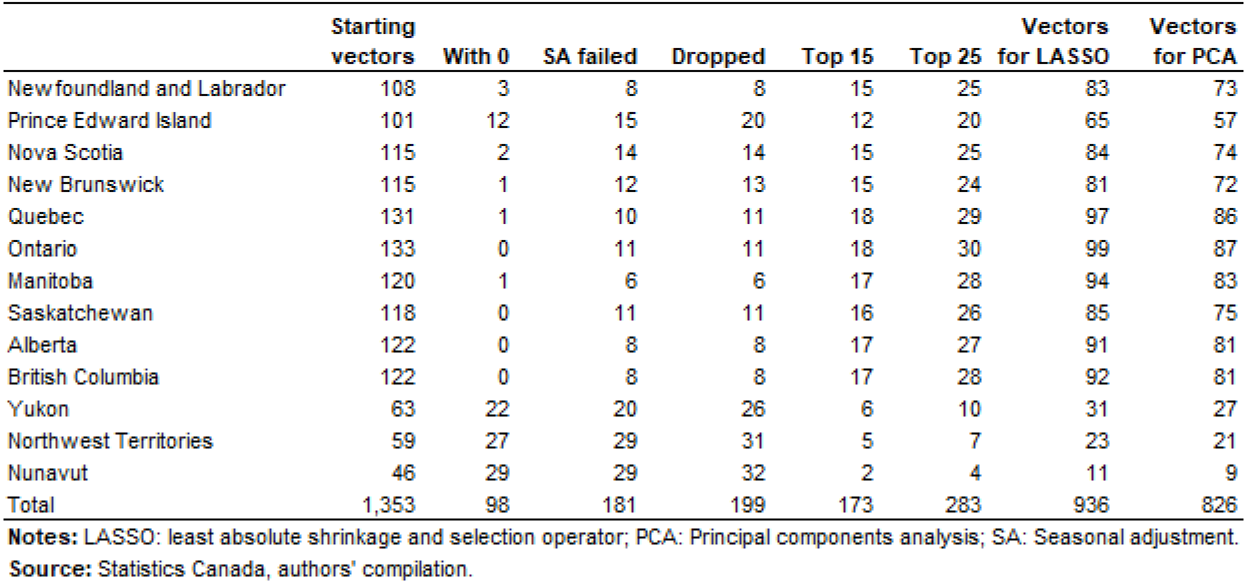
The function $f(.)$ is a set of instructions for transforming a large number inputs into a single series. In this paper, it is assumed that the function can be approximated based on a linear combination of inputs, and that the inputs can be selected either by selecting a subset of the available data or by creating a combination of all input data series. Below, three approaches are explored: 1) a simple model; 2) PCA; and 3) LASSO. The simple and the LASSO model fall in the former category while PCA falls in the latter category.

The assumptions and implementation of the models is discussed in detail below. Across modelling strategies, the following steps are followed to estimate index values in all cases:

- 1: Prepare the input data.

The input data set has 1,341 series that are distributed unevenly across the provinces and territories (Table 3). Not all series have equal utility for modelling aggregate economic activity. In cases where seasonal adjustment failed, the series are removed. Similarly, series with 0 values are removed. These tend to be series where 0 values are interspersed with nominal values. In these cases, seasonally adjusted values can be negative, growth rate or log-difference transformations do not work, and the series have questionable value for use as an ongoing indicator of economic activity. Overall, 198 variables are dropped for these reasons.

Table 3
Number of input variables



	Starting vectors	With 0	SA failed	Dropped	Top 15	Top 25	Vectors for LASSO	Vectors for PCA
Newfoundland and Labrador	108	3	8	8	15	25	83	73
Prince Edward Island	101	12	15	20	12	20	65	57
Nova Scotia	115	2	14	14	15	25	84	74
New Brunswick	115	1	12	13	15	24	81	72
Quebec	131	1	10	11	18	29	97	86
Ontario	133	0	11	11	18	30	99	87
Manitoba	120	1	6	6	17	28	94	83
Saskatchewan	118	0	11	11	16	26	85	75
Alberta	122	0	8	8	17	27	91	81
British Columbia	122	0	8	8	17	28	92	81
Yukon	63	22	20	26	6	10	31	27
Northwest Territories	59	27	29	31	5	7	23	21
Nunavut	46	29	29	32	2	4	11	9
Total	1,353	98	181	199	173	283	936	826

Notes: LASSO: least absolute shrinkage and selection operator; PCA: Principal components analysis; SA: Seasonal adjustment.
Source: Statistics Canada, authors' compilation.

The filtered input data set then contains 1,143 series. However, the series are typically reported in levels (e.g. hours worked or manufacturing sales in chained dollars) and present strong trends over the sample period. To account for the trends, the series are transformed to month-to-month growth rates. These growth rates will ultimately be compared to measures of real GDP growth, and they have the advantage of being bound by -100% for the maximum decline.[4] The growth rates for the series are assumed to be covariance stationary.

To use the series in estimation, all series are demeaned and scaled to have unit variance at a monthly frequency. This normalization process is applied to variables to prevent variables with naturally larger unit values from affecting results. Because the monthly time series can be have high variability, and because periods of economic shocks such as recessions can produce aberrant data points, all series are winsorized (or top and bottom coded) prior to estimation. This prevents extreme data points from affecting results. For creating monthly indexes, the parameter values from models based on the winsorized data are combined with un-winsorized data which permits larger values to have their full influence when large shocks occur.

Finally, the noisiest series are removed. For PCA, the top 25% of series by variance are removed by province while for LASSO the top 15% of series by variance are removed. These thresholds are arbitrary, but their imposition was found to improve the ability of the models to inform on aggregate activity (i.e. improve the signal relative to noise), and to improve consistency of results across methods. After adjusting for high variance series there are a total of 928 input variables for LASSO and 820 for PCA. Nunavut has the fewest series available while Ontario and Quebec have the most.

[4] It is more traditional to use log-differences in time series analysis. Here, growth rates are used because the rate of change is bounded by -100%, which is not the case with log-differences. Implementing a proper review of the stationarity properties of the input series remains a priority for future work.

- 2: Using the winsorized growth rates, calculate the annual average of monthly growth rates for use in the models.
- 3: Estimate the model parameters.

The estimation is initiated by combining real GDP growth with annual averages of monthly input series for years 2002 to 2018. Real GDP growth is not scaled or demeaned. Using real GDP growth as the target variable and the annual averages of monthly series as input variables, the functional form of $f\left(.\right)$ is estimated. In all models employed, it is assumed that a linear combination of input variables can be used to transform the multiplicity of input variables into a single measure of aggregate activity. In the case of the simple model and LASSO, a subset of the variables is used directly. In the case of PCA, the first ten principal components are employed as the starting point. It is also assumed that OLS can be used to generate contributions for combining input variables. By using regression methods to combine inputs, a further assumption that economic structures are, on average, the same over the entire sample period is imposed.

- 4: Use the model to estimate monthly growth rates.

Since all approaches can be viewed as an OLS regression with demeaned inputs, the intercept can be interpreted as the average annual growth rate of real GDP between 2002 and 2018. The selected inputs (monthly time series or principal components) produce fluctuations around this average growth rate. For 2019 and 2020, it is assumed that the average growth rate from 2002 to 2018 is representative of underlying growth.

To produce monthly estimates, it is necessary to adjust parameter estimates or monthly series to account for the difference in periodicity. The model constant is adjusted to a monthly frequency based on the monthly compound growth rate that is equivalent to the annual estimate:

$$\hat{\beta}_{0,monthly} = (\hat{\beta}_{0,annual} + 1)^{1/12} - 1$$

To estimate monthly fluctuations around the trend growth rate, the raw input series are used. This allows large fluctuations in the time series to present their full impact when economic shocks, such as recessions or commodity price cycles, impact provincial and territorial economies. The monthly inputs have their variances adjusted to match the annual variance prior to use. OLS regression estimates are based on the ratio of cov(x,y)/var(x). In the current context, aggregation through time reduces the variance of the X matrix. To account for this, the variance of the monthly data is re-scaled to match the variance of the annual data for each series as:

$$\sigma_{monthly,adjusted} = \left(\frac{\left(x - \mu_x\right)}{\sigma_{x,monthly}} \right) * \sigma_{x,annual} + \mu_x$$

- 5: Generate level indexes

The fitted values from the models are estimates for monthly growth in economic activity. They can be transformed into indexes by adding 1 to create a linking value for a chain index. The index level is then calculated by chaining forward from January 2002.

The growth rate estimates have a confidence interval associated with them as there is quantifiable uncertainty that comes from the model. There is also unquantifiable uncertainty that arises from possible model misspecification. To produce a level index, it is necessary to assume that the growth rate estimates are sufficiently accurate that they can be employed for chaining even though errors are compounded through time. This is a strong assumption, but is consistent with the way mean values from survey data for prices, values and quantities are combined to produce chain-quantity or chain-price indexes.

The use of multiple models in step 3 will ultimately lead to different flavors of the activity index being presented. In the current context, where statistical models are being used to inform about economic activity in an environment where they cannot be optimally implemented, the creation of multiple versions of the activity index serves an important role. Because the true value for $y_{monthly,t}$ is unknown, validating $f(.)$ and $\hat{y}_{monthly,t}$ is challenging. The outputs from the different methods provide a natural form of data confrontation which helps to gauge the adequacy and generalizability of the estimates.

3.1.1 Simple model

The simple model imposes the a priori assumption that total employment, total exports and total retail sales contain the appropriate information for understanding aggregate economic fluctuations. This is likely too strong an assumption as more than three inputs are needed to fully capture the complexities of aggregate economic activity. However, the model is consistent across all provinces and territories, and it is straightforward to understand. It, therefore, has value as a base against which more complex methods can be assessed. The simple approach also represents a method that can be viewed as consistent with the types of projectors that are used to infer gross output movements that are used as inputs for monthly GDP for Canada (Statistics Canada 2020b).

Since the series do not have a natural aggregation structure for combining them, regressions are used to determine relative contributions of the variables rather than an index number formula. Because the inputs are assumed to be the necessary inputs, the three series are included regardless of their statistical significance in regressions.

3.1.2 Principal components analysis

PCA is a variable reduction technique that aims to explain the variance of a given data set using a smaller number of principal components (OECD 2008, Jollife 2002). A data set with p variables $X = \left[x_1, x_1, ..., x_p\right]$, can be transformed to produce p principal components:

$$Z = AX$$

where

$$z_1 = a_{1,1}x_1 + a_{1,2}x_2 + .. + a_{1,p}x_p$$

$$z_2 = a_{2,1}x_1 + a_{2,2}x_2 + .. + a_{2,p}x_p$$

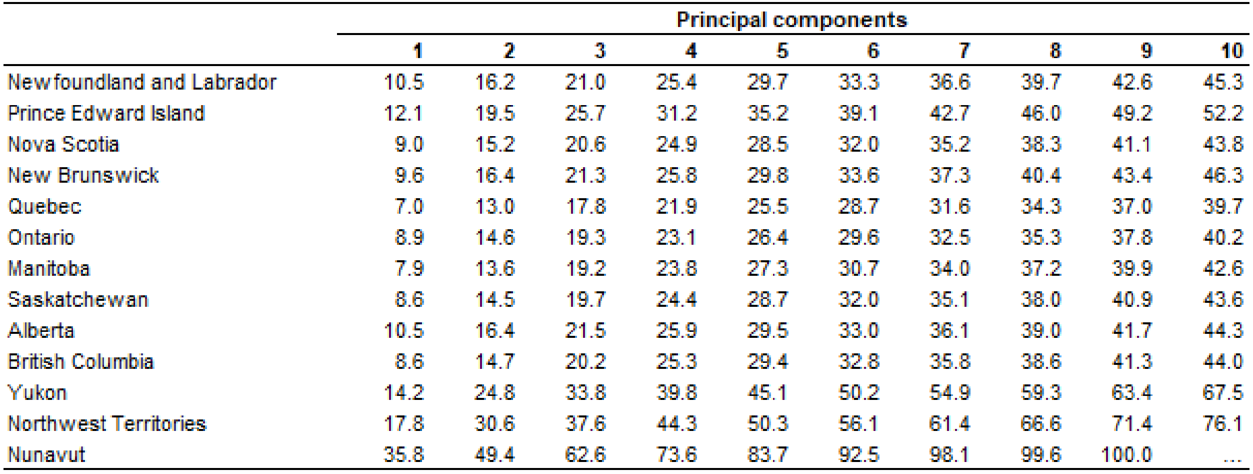
$$z_p = a_{p,1}x_1 + a_{p,2}x_2 + .. + a_{p,p}x_p$$

The principal components are constructed as linear combinations of the input variables (the monthly series). The first principal component explains the largest proportion of the variance of the input variables. It is the eigenvector associated with the largest eigenvalue. The second principal component is orthogonal (uncorrelated) with the first principal component and is the eigenvector associated with the second largest eigenvalue. It explains the second largest component of the input variables. It can, therefore, be said to measure a different statistical dimension of the available series. The third principal component is orthogonal to the first two and measures the third largest proportion of variance in the data. And so on until the p^{th} principal component.

To implement PCA here, the principal components that are used for model estimation and the loadings are determined using the winsorized, demeaned, unit variance monthly time series. The loadings are then applied to the raw, un-winsorized series to produce the raw principal components that are used to predict the monthly growth rates.

When PCA works well, a large portion of the variance of a data set can be explained by the first few principal components, and only the first few principal components are used for analysis. Unfortunately, in the case of the input data set for the provincial and territorial economies, PCA does not work well for reducing the scope of information in the data set (Table 4). The first principal component typically accounts for less than 10% of the variation in the data set. And, when averaged to produce an annual frequency estimate, the first principal component does not correlate well with annual real GDP growth for most provinces and territories (Table 5).

Table 4
Percent of variation by principle component



	Principal components									
	1	2	3	4	5	6	7	8	9	10
Newfoundland and Labrador	10.5	16.2	21.0	25.4	29.7	33.3	36.6	39.7	42.6	45.3
Prince Edward Island	12.1	19.5	25.7	31.2	35.2	39.1	42.7	46.0	49.2	52.2
Nova Scotia	9.0	15.2	20.6	24.9	28.5	32.0	35.2	38.3	41.1	43.8
New Brunswick	9.6	16.4	21.3	25.8	29.8	33.6	37.3	40.4	43.4	46.3
Quebec	7.0	13.0	17.8	21.9	25.5	28.7	31.6	34.3	37.0	39.7
Ontario	8.9	14.6	19.3	23.1	26.4	29.6	32.5	35.3	37.8	40.2
Manitoba	7.9	13.6	19.2	23.8	27.3	30.7	34.0	37.2	39.9	42.6
Saskatchewan	8.6	14.5	19.7	24.4	28.7	32.0	35.1	38.0	40.9	43.6
Alberta	10.5	16.4	21.5	25.9	29.5	33.0	36.1	39.0	41.7	44.3
British Columbia	8.6	14.7	20.2	25.3	29.4	32.8	35.8	38.6	41.3	44.0
Yukon	14.2	24.8	33.8	39.8	45.1	50.2	54.9	59.3	63.4	67.5
Northwest Territories	17.8	30.6	37.6	44.3	50.3	56.1	61.4	66.6	71.4	76.1
Nunavut	35.8	49.4	62.6	73.6	83.7	92.5	98.1	99.6	100.0	...

... not applicable

Source: Statistics Canada, authors' compilation.

The correlations indicate that outside of Alberta, British Columbia and Ontario, using only the first principal component will not produce an activity index that provides a suitable measure for determining the performance of the provinces and territories based on aggregating month-to-month fluctuations. As a result, an activity index based only on the first principal component, such as the one produced by the Federal Reserve Board of Chicago (Federal Reserve Board of Chicago 2020, Brave and Butters 2010, Evans and Pham-Kanter 2002), is not pursued here.

Table 5
Correlation between first principle component and annual real gross domestic product (GDP) growth

	Winsorized index	Un-Winsorized index
Newfoundland and Labrador	-0.349	-0.256
Prince Edward Island	-0.199	-0.307
Nova Scotia	0.531	0.491
New Brunswick	-0.709	-0.673
Quebec	0.657	0.661
Ontario	0.775	0.778
Manitoba	0.562	0.567
Saskatchewan	0.477	0.338
Alberta	0.937	0.940
British Columbia	0.840	0.848
Yukon	0.177	0.251
Northwest Territories	0.423	0.422
Nunavut	0.355	0.444

Source: Statistics Canada, authors' compilation.

While the first principal component has difficulties correlating with annual fluctuations in real GDP, this does not mean there is no information in the first few principal components for explaining real GDP growth. Therefore, to generate a model based on the principal components, regressions are performed on all combinations of the first 10 principal components as regressors for explaining real GDP growth. The regression that maximizes the adjusted R-squared is then selected as the preferred model. This produces 13 models that perform reasonably well for explaining real GDP growth. Moreover, the models generally do well for explaining the 2008 recession and other, province- and territory-specific fluctuations.

3.1.3 Least Absolute Shrinkage and Selection Operator

LASSO is the solution to a constrained optimization problem similar to OLS. Under classical linear regression, $X = \left[x_1, x_2, ..., x_p \right]$ is a n x p matrix holding the predictor variables which are used to explain the variation in a target vector y of length n. The coefficients for the regression $\hat{\beta} = \left[\beta_0, ..., \beta_p \right]$ are then the solution to the problem that seeks to minimization the sum of the squared errors between y and a linear combination of the variables in X :

$$\hat{\beta}_{OLS} \mid= \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - ... - \beta_p x_{ip})^2$$

LASSO (Tibshirani 1996) is one of a class of estimators that seeks to penalize the OLS estimator for over fitting (i.e. including too many variables) through its regulatory parameter λ. It is similar to using an adjusted R-squared or information criterion to penalize for including too many regressors. However, it goes further than penalizing for extra regressors when looking at model quality. It selects relevant variables. LASSO is the solution to:

$$\hat{\beta}_{LASSO} \mid= \underset{\beta}{\operatorname{argmin}} \left[\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - ... - \beta_p x_{ip})^2 + \lambda \sum_{j=1}^{p} (\left| \beta_j \right|) \right]$$

The $\lambda \geq 0$ parameter controls the strength of the penalty, the larger the value of lambda, the greater the amount of shrinkage. The LASSO algorithm is only permitted to include values for β_j

up to a particular absolute total. As a result, LASSO sets less consequential variable coefficients to 0. This can be viewed as similar to the type ore result found using a general-to-specific modelling strategy, but which is applicable on a larger scale. The result is a method for dealing with large data sets where a large number of predictors can be included, and the algorithm will select those whose covariance properties are most important for predicting the target variable (y).

LASSO comes with its own limitations. In cases when groups of predictor variables are highly correlated with each other, LASSO tend to keep one variable from each group and shrink the coefficient of the other variables to zero. And in other cases, when the data set has small n and large p, LASSO selects at most n variables before it is saturated. However, there may be more than n variables with non-zero coefficient in the true model.

The Elastic Net method (Zou and Hastie 2005) is an extension of LASSO. By controlling the penalty weight α, the Elastic Net model stabilizes the variable selection from a group of correlated variables and removes the limitation on the number of variables selected. The coefficients are estimated as follows:

$$\hat{\beta}_{EN}\text{À} = \underset{\beta}{\operatorname{argmin}}\left\{\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_p x_{ip})^2 + \lambda\sum_{j=1}^{p}\left[\frac{1}{2}\left(1-\alpha\right)\beta_j^2 + \alpha\left|\beta_j\right|\right]\right\}$$

Where $0 \leq \alpha \leq 1$ is the penalty weight. With α equal to 1, the Elastic Net is the same as the LASSO model and, with α close to 1, the Elastic Net behave similar to LASSO, but removes the problematic behavior caused by high correlations among variables.

The outputs of LASSO in terms of the number of variable selected and their statistical significant were carefully studied. In almost all cases where LASSO worked, LASSO seems to include variables in the model that are not statistical significant. To ensure that the relation between the target y and regressors is justifiable with a better statistical result, a step wise regression with backward selection is used on the variables selected by LASSO to remove non-significant variables from the model.

Step wise regression is a method that examines the statistical significant of each independent variable within the model. It builds a model by successively adding (forward selection) or removing (backward selection) variables based on the t-statistics of their estimated coefficients. The backward elimination method begins by including all variables in the model, then each variable is removed one at a time, to test its importance. Those variables that are not statistically significant are removed from the model.

The LASSO model did not select any variables for New Brunswick, Nova Scotia, Ontario and Northwest territories. The Elastic Net method is for used for these jurisdictions instead. And for the other two territories, Yukon and Nunavut, a manual step wise regression is performed.

In both methods, LASSO and Elastic Net, cross validation from the R package caret is used to tune parameters lambda and alpha. The cross validation uses a rolling forecasting origin technique (Hundman and Athanasopoulos 2014) instead of the simple random sampling. This technique is specific to time series data sets.

4 Monthly index assessment

The three approaches have different strengths and weaknesses, which affects their use (Table 6). The simple index and the LASSO index have the strength that their models are parsimonious, and the indexes they produce are less noisy than PCA-based indexes. However, these indexes are based on a greatly reduced set of variables, which for the simple indexes are often statistically insignificant in annual regressions. These indexes also tend to focus on employment series rather than a broad range of economic activities, and so may not present ideal predictors of monthly activity fluctuations if changes in production are not contemporaneously aligned with labour variables.

The PCA index has the strength that the methodology is sound and well understood. It works for all provinces and territories. However, it produces the noisiest activity indexes making them difficult to interpret, and in some cases (e.g., NL) the index can decline sharply. The PCA indexes are also combined based on maximizing the adjusted R-squared across regressions. This produces a linear combination of principal components that are statistically significant and insignificant. These inclusions err on the side of adding additional information that includes some noise as it is not clear that data at an annual frequency represents month-to-month variability.

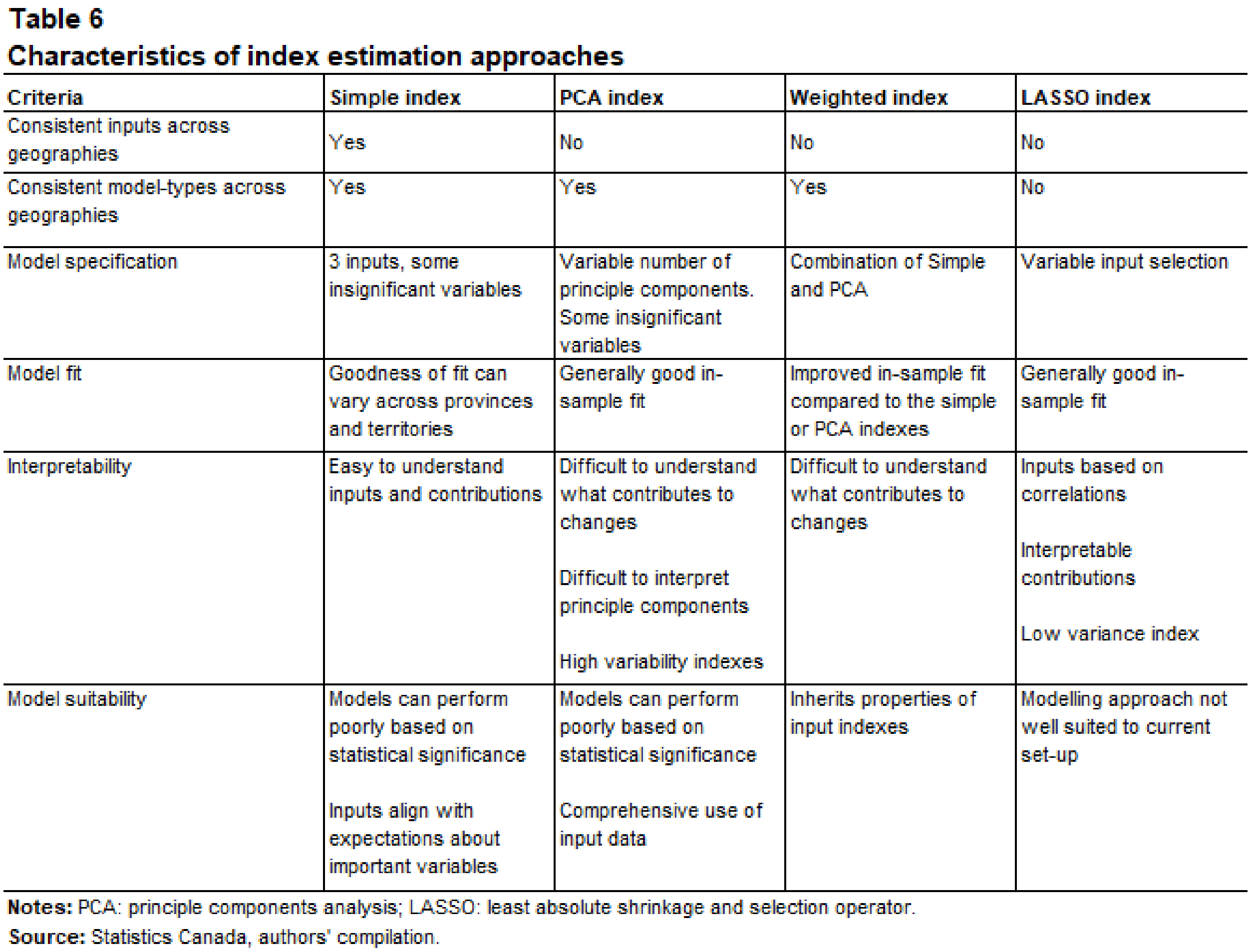


Table 6
Characteristics of index estimation approaches

Criteria	Simple index	PCA index	Weighted index	LASSO index
Consistent inputs across geographies	Yes	No	No	No
Consistent model-types across geographies	Yes	Yes	Yes	No
Model specification	3 inputs, some insignificant variables	Variable number of principle components. Some insignificant variables	Combination of Simple and PCA	Variable input selection
Model fit	Goodness of fit can vary across provinces and territories	Generally good in-sample fit	Improved in-sample fit compared to the simple or PCA indexes	Generally good in-sample fit
Interpretability	Easy to understand inputs and contributions	Difficult to understand what contributes to changes Difficult to interpret principle components High variability indexes	Difficult to understand what contributes to changes	Inputs based on correlations Interpretable contributions Low variance index
Model suitability	Models can perform poorly based on statistical significance Inputs align with expectations about important variables	Models can perform poorly based on statistical significance Comprehensive use of input data	Inherits properties of input indexes	Modelling approach not well suited to current set-up

Notes: PCA: principle components analysis; LASSO: least absolute shrinkage and selection operator.
Source: Statistics Canada, authors' compilation.

Combining the indexes provides an additional method for their use. Since the simple index is relatively stable, but focuses on a limited number of fundamental series, and the PCA is more variable, but includes linear combinations of all inputs, these series are combined to produce a weighted index that has better characteristics than the components. As with the regression coefficients, annual real GDP growth is used as the comparison as it is the primary source of aggregate economic activity available for the provinces and territories. To combine the PCA index and simple index, values of nu between 1% and 100% are used to create weighted indexes as:

$$weighted\_index = (1-\nu) * simple\_index + \nu * activity\_index$$

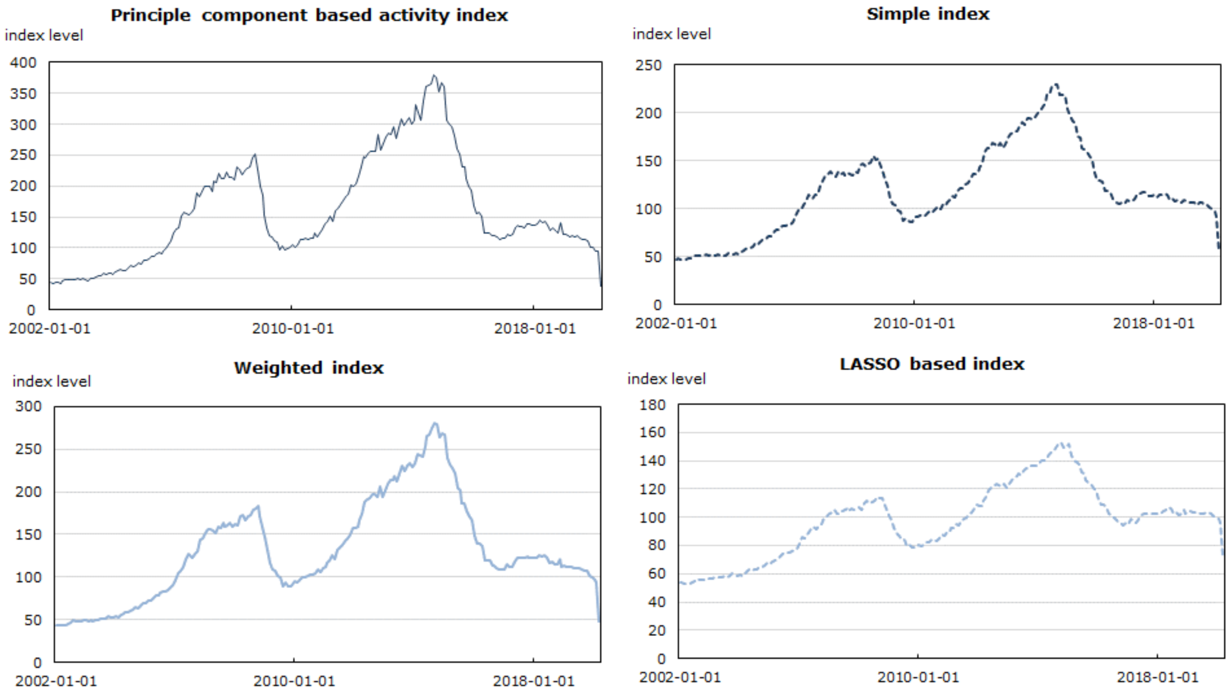
The nu corresponding to the weighted index that has the highest correlation with real GDP growth is then selected.

The methods for generating monthly indexes generally return similar types of information on economic cycles and major economic shocks in the provinces and territories. As examples, the indexes for Alberta (Panel 1) and Newfoundland and Labrador (Panel 2), are presented below.

For larger economics, such as Alberta, all approaches return similar information on periods of growth or decline, but the magnitude of the cycles can differ depending on methodology. In general, the PCA based indexes have the largest variability while the simple index has the least. In some cases, such as the PCA index for Newfoundland and Labrador, the model fails to produce a reasonable result. In these cases, the index will not be made available, and is deemed not-fit-for-use. Nevertheless, when the indexes appear to have the appropriate characteristics, there is a strong correlation across measures for the implied economic activity, and the movement of the indexes through time corresponds with what is known about provincial and territorial economic performance.

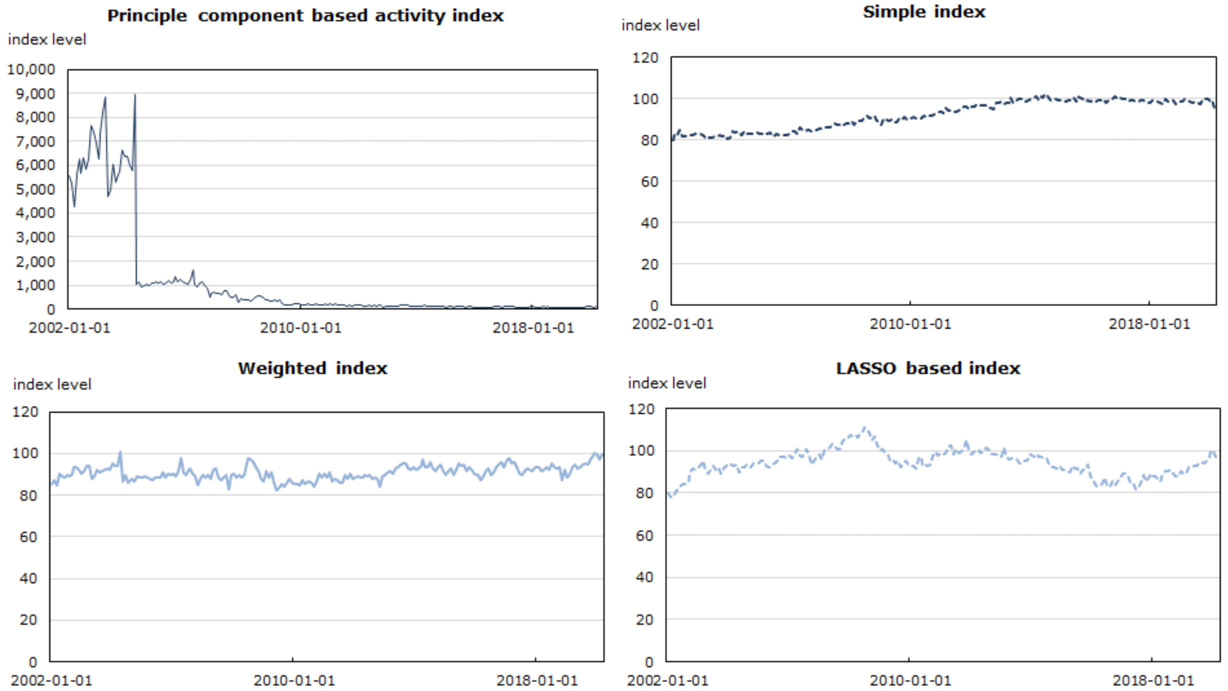
Additionally, comparisons with sub-annual real GDP estimates for Ontario and Quebec show that the year-to-year growth rates are highly correlated but that business cycles can be accentuated in the activity indexes. The indexes, therefore, appear to capture relevant information for economic cycles, periods of stronger or weaker growth and for understanding economic performance. They do not, however, correspond directly with real GDP, and should not be interpreted as a direct measure of real GDP.

Chart 1
Alberta: activity indexes; nu=0.39



Notes: LASSO = Least absolute shrinkage and selection operator.
Source: Statistics Canada, authors' calculations.

Chart 2
Newfoundland and Labrador: activity indexes; nu=0.29


Principle component based activity index

Simple index

Weighted index

LASSO based index

Notes: LASSO = Least absolute shrinkage and selection operator.
Source: Statistics Canada, authors' calculations.

5 Conclusion

Measures of aggregate economic activity for economies are important for informing decisions about fiscal and monetary policy, for determining the characteristics of business cycles and for examining economic performance. In this study, four indexes of provincial and territorial economic activity based on different methodological approaches are estimated and presented. The methodologies are based on a simple model; PCA; a weighted combination of the simple index and the PCA index; and LASSO. In most cases, all approaches produce roughly similar results. However, the degree of cyclicality and the variance of month-to-month changes can differ significantly. As a general rule, PCA produces the greatest variability and the largest cycles while the simple index is the most stable.

Based on the properties of the methodologies and their outputs, the simple index is the most consistent across provinces and territories. It is also the easiest to interpret in terms of variable contributions and justification for variable inclusion. However, parameter values are often statistically insignificant and the input series are chosen as much for their economic importance as for their presence in all jurisdictions. As a result, these models offer a more limited approach for examining aggregate economic activity, but also present a basis for comparisons to more complex models.

Indexes based on PCA appear to offer a more complete sense of how activity is evolving over time, but it is unclear at the moment how the principal components should be interpreted. Because of this, and because the PCA indexes have the largest variability, they present a trade-off between overall use of input series and interpretability.

Weighting the simple index and PCA index produces a result that has a superior correlation with annual GDP fluctuations. The weighted combination continues to have more variability than the simple index. Since the PCA is included, it is also not as easy to interpret as the simple index, but likely provides a better measure of aggregate activity than its constituent parts.

The LASSO index performs well when compared to annual real GDP, but the model set-up is not as well suited to the situation encountered when trying to estimate the activity indexes. In particular, the relatively small number of observations limits the ability of the algorithms to perform cross validation. Moreover, while the input series are a distinct subset of the input data set, and their contributions can be generated in a straightforward manner, there is no theoretical reason for why the variables are important, and this limits the model's interpretability.

Given the strengths and weaknesses present between the suitability of the models, their performance and examinations of their outputs, the assessments made thus far suggest that the simple indexes or LASSO indexes present results related to a set of fundamental inputs (often heavily influenced by employment series), that the PCA indexes relate more to some form of short-term activity (but the signal is noisy), and that the weighted index presents a compromise between the two.

The indexes as currently estimated are correlated with annual measures of real GDP and sub-annual measures of real GDP for Ontario and Quebec, but they should not be interpreted as being a real GDP measure. The indexes display greater variability and cyclicality that real GDP measures, and are constituted from measures of gross outputs, employment, relative prices and important ratios such as the unemployment rate. This makes the indexes appropriate for understanding economic activity, but they are not real GDP. Moreover, the indexes do not inform about differing levels of economic activity between provinces and territories.

The indexes are also based on an input data set and modelling strategies that are not ideal. Numerous assumptions must be imposed to produce the indexes, any of which may be a source of important measurement errors. As a consequence, the indexes presented here should be viewed as experimental, and are subject to revision or replacement as future research improves the processes and/or test the assumptions for their validity.

At the current time, the correlations between the different approaches, their positive correlation with provincially produced measures of sub-annual real GDP and examinations of their properties against known provincial and territorial economic performance supports their use as indicators of business cycles, for understanding the magnitude of shocks relative to a provinces' or territory's history and for understanding how regional economies are progressing. Inter-provincial comparisons are also supported, but with the caveat that model performance is difficult to understand in all situations, and that level comparisons across provinces are not possible using the index values.

References

Brave, Scott., and R. Andrew Butters. 2010. “Chicago Fed National Activity Index Turns Ten—Analyzing Its First Decade of Performance.” *Chicago Fed Letter*, no. 273 (April). Federal Reserve Bank of Chicago. https://www.chicagofed.org/~/media/publications/chicago-fed-letter/2010/cflapril2010-273-pdf.pdf.

Statistics Canada 2020a.” Gross domestic product (GDP) at basic prices, by industry, monthly (36100434).” Statistics Canada. Statistics Canada, January 24, 2020. https://www150.statcan.gc.ca/n1/pub/13-607-x/2016001/230-eng.htm (accessed June 2, 2020)

Statistics Canada 2020b.” Gross domestic product (GDP) at basic prices, by industry, monthly (36100434).” Statistics Canada. Statistics Canada, July 31, 2019. https://www.statcan.gc.ca/eng/statistical-programs/document/1301_D1_V3 (accessed June 2, 2020)

Federal Reserve Board of Chicago 2020. “Chicago Fed National Activity Index (CFNAI) Current Data.” Federal Reserve Board of Chicago, June 22, 2020. https://www.chicagofed.org/research/data/cfnai/current-data (accessed June 2, 2020).

Evans, Liu, Charles L., and Genevieve Pham-Kanter. 2002. “The 2001 Recession and the Chicago Fed National Activity Index: Identifying Business Cycle Turning Points.” *Economic Perspectives* 26 (3). Federal Reserve Bank of Chicago: 26–43. https://www.chicagofed.org/~/media/publications/economic-perspectives/2002/3qepart2-pdf.pdf.

Hyndman, R.J., & Athanasopoulos, G. (2018) *Forecasting: principles and practice*, 2nd edition, OTexts: Melbourne, Australia. OTexts.com/fpp2. Accessed on June 2, 2020

Jollife, I.T. 2002. *Principle Components Analysis Second Edition*. Springer-Verlag New York Inc, New York, NY.

United Nations (UN), European Commission (EC). International Monetary Fund (IMF), Organisation for Economic Co-operation and Development (OECD), and World Bank (WB). 2009. *System of National Accounts, 2008.* New York: United Nations. Available at: https://unstats.un.org/unsd/nationalaccount/docs/sna2008.pdf (accessed June 2, 2020).

Organisation for Economic Co-operation and Development (OECD). 2008. *Handbook on Constructing Composite Indicators Methodology and User Guide*. Organization for Economic Development. https://www.oecd.org/sdd/42495745.pdf (accessed June 2, 2020).

Tibshirani, Robert. 1996. “Regularization Shrinkage and Selection via the Lasso.” *Journal of Royal Statistical Society: Series B*.

Zou, Hui, and Trevor Hastie. 2005. “Regularization and Variable Selection via the Elastic Net.” *Journal of Royal Statistical Society: Series B (Statistical Methodology)* 67 (2): 301–20. doi:10.1111/j.1467-9868.2005.00503.x.