SDMXUSE

MODULE TO IMPORT DATA FROM STATISTICAL AGENCIES USING THE SDMX STANDARD

UCL Université catholique de Louvain

Sébastien Fontenay

sebastien.fontenay@uclouvain.be

Motivation

- Nowcasting Euro Area GDP
  - i.e. computing early estimates of current quarter GDP
    - because official estimates are published with a considerable delay (e.g. Eurostat flash estimate is released 6 weeks after the end of the quarter)
- Statistical models can perform this exercise by exploiting more timely information
  - Financial series
    - E.g. market indices, commodity prices, interest rates
  - Business & consumer surveys
    - E.g. EU harmonised surveys, Economic Sentiment Indicator, Markit PMI
  - Real activity series
    - E.g. industrial production index or retail sales

MOTIVATION

- Mixed-frequency problem
  - This timely information has monthly or higher frequency while GDP is quarterly
- Traditional method to deal with this: bridge equations
  - Regression of quarterly GDP growth on a small set of preselected key monthly indicators
    - Usually a few predictor variables (hand-selected or using variables selection methods – e.g. Lasso) considered in terms of quarterly averages
      - One issue is that it requires forecasting any months of current quarter for which data is not yet available (ragged edge problem)
- Special "bridging" technique: blocking approach
  - Following Carriero *et al.* (2012), we split the high frequency information into multiple low frequency time series
    - We will therefore obtain 3 quarterly series for a given monthly variable
      - Better at dealing with ragged edge problem, as we use only actual monthly observations that are available for the quarter

MOTIVATION

Consumer confidence indicator EA19	
Jan-2016	- 6.3
Feb-2016	- 8.8
Mar-2016	- 9.7
Apr-2016	- 9.3
May-2016	- 7
Jun-2016	- 7.2
Jul-2016	- 7.9
Aug-2016	- 8.5
Sep-2016	N/A

- The first quarterly series (M1) collects observations from the first months of each quarter (i.e. January, April, July and October)
- The second one (M2) collects observations from the second months (i.e. February, May, August and November)
- The last one (M3) assembles the observations from the third months (i.e. March, June, September and December)

	M1	M2	M3
Q1	- 6.3	- 8.8	- 9.7
Q2	- 9.3	- 7	- 7.2
Q3	- 7.9	- 8.5	N/A

```
. sdmxuse data ESTAT, dataset(ei_bsco_m) dimensions(.BS-CSMCI.SA..EA19) start(2016)
```

MOTIVATION

Example of Stata code to implement the blocking approach

```
. sdmxuse data ESTAT, dataset(ei_bsco_m) dimensions(.BS-CSMCI.SA..EA19) start(2016)
. keep time value
. gen time2 = month(dofm(monthly(time, "YM")))
. tostring time2, replace
. replace time2="M1" if inlist(time2, "1", "4", "7", "10")
. replace time2="M2" if inlist(time2, "2", "5", "8", "11")
. replace time2="M3" if inlist(time2, "3", "6", "9", "12")
. reshape wide value, i(time) j(time2, string)
. gen time2=qofd(dofm(monthly(time, "YM")))
. drop time
. rename time2 time
. collapse valueM1 valueM2 valueM3, by(time)
. tsset time, quarterly
```

MOTIVATION

- Another problem is that "the number of candidate predictor series (N) can be very large, often larger than the number of time series observations (T)" leading to a so-called high-dimensional problem (Stock & Watson, 2002)
  - In order to exploit all the information, Stock & Watson (2002) propose to model the covariability of the predictor series in terms of a relatively few number of unobserved latent factors
    - They estimate the factors using principal components and show that these estimates are consistent in an approximate factor model even when idiosyncratic errors are serially and cross-sectionally correlated
      - Recent works have shown that regressions on factors extracted from a large panel of time series outperform traditional bridge equations (e.g. Barhoumi *et al.*, 2008)

MOTIVATION

- The estimation is carried out in two steps:
  - First, the factor analysis shrinks the vast amount of information into a limited set of components:

$$X_t = \Lambda F_t + e_t \qquad (1)$$

    - with X_t a N-dimensional multiple time series of candidate predictors, F_t a K-dimensional multiple time series of latent factors (with $K < N_t$), Λ a matrix of loadings relating the factors to the observed time series and e_t are idiosyncratic disturbances

  - Second, the relationship between the variable to be forecast and the factors is estimated by a linear regression:

$$y_t = c + \alpha\, w_t + \sum_{j=1}^{K} \beta_j\, f_{jt} + \varepsilon_t \qquad (2)$$

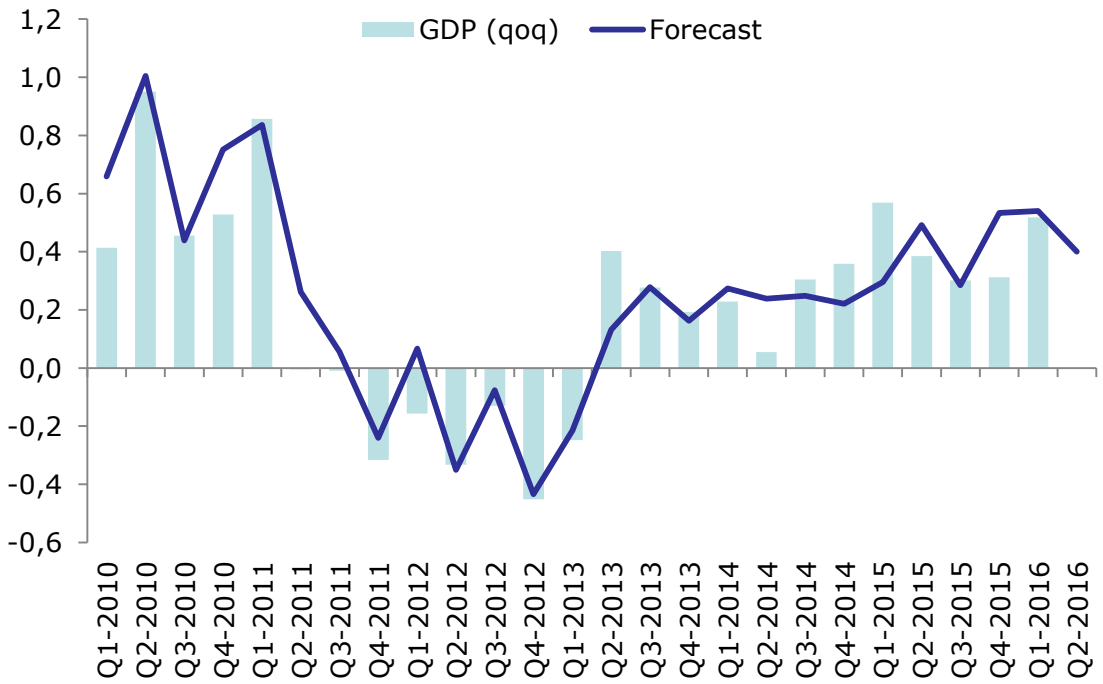
    - with y_t the log-difference of the quarterly GDP, w a vector of observed variables (e.g. lags of y), f_{jt} the K factors identified above and ε_t the resulting forecast error

MOTIVATION

Pseudo out-of-sample evaluation

- We replicate the data availability of monthly time series by estimating the model for each period using only the information available at the end of the reference quarter
  - E.g. only first month for industrial production index and retail sales, two first months for unemployment indicators and all three months for survey data

Mean Absolute Error	0,11
Root Mean Squared Error	0,14



Motivation

- But how do we get these time series (often more than one hundred) updated immediately after new releases are made available?
  - Objective is to run forecasting model every time new data is made available to observe changes in the prediction
    - At the beginning of the quarter, only financial series are available but they are weakly correlated with GDP
    - At the end of each month, business and consumer surveys are available and bring some valuable insights on the current economic situation
    - Towards the end of the reference quarter, real activity series (notably production indices) for the first month of the quarter become available; usually associated with GDP volatility

Motivation

September 2016						
Mon	Tues	Wed	Thurs	Fri	Sat	Sun
29	30 ESTAT – B&C surveys	31 ESTAT – Unemployment	1	2	3	4
5 ESTAT – Serv. turnover	6 ESTAT – GDP	7	8 OECD – Lead. indicators	9	10	11
12 ECB – Interest rates	13 ESTAT – Employment	14 ESTAT – Indus. production	15 ESTAT – HICP	16 ECB – Car registrations	17	18
19	20	21	22 ESTAT – Flash consumer conf.	23	24	25
26	27 ECB – Monet. aggregates	28	29 ESTAT – B&C surveys	30 ESTAT – Unemployment	1	2

SDMX STANDARD

- SDMX stands for Statistical Data and Metadata Exchange
  - Initiative started in 2001 by 7 international organisations
    - Bank for International Settlements (BIS), European Central Bank (ECB), Eurostat (ESTAT), International Monetary Fund (IMF), Organisation for economic Co-operation and Development (OECD), United Nations (UN) and the World Bank (WB)
  - Their objective was to develop more efficient processes for sharing of statistical data and metadata
    - Metadata = data that provides information about other data
      - e.g. the data point 9.9 is not useful without the information that it is a measure of the total unemployment rate (according to ILO definition) for France, after seasonal adjustment but no calendar adjustment, in June 2016

SDMX STANDARD

- The initiative evolved around two axes:
  - setting technical standards
    - for compiling statistical data
      - the SDMX format (built around XML syntax) was created for this purpose
  - and developing statistical guidelines
    - i.e. a common metadata vocabulary to make international comparisons meaningful
- The primary goal was to foster data sharing between participating organisations using a “pull” rather than a “push” reporting format
  - i.e. instead of sending formatted databases to each others, statistical agencies could directly pull data from another provider website
    - For this purpose, they created RESTful web services

SDMX STANDARD

- Concretely, users can access a dataset (when they know its identifier) by sending an HTTP request to the URL of the service
  - The result is a structured (SDMX-ML) file
    - E.g. http://ec.europa.eu/eurostat/SDMX/diss-web/rest/data/teilm020/all?

```
▼<message:GenericData xmlns:footer="http://www.sdmx.org/resources/sdmxml/schemas/v2_1/message/footer"
 xmlns:common="http://www.sdmx.org/resources/sdmxml/schemas/v2_1/common" xmlns:message="http://www.sdmx
 xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
  ▶<message:Header>...</message:Header>
  ▼<message:DataSet structureRef="ESTAT_DSD_teilm020_1_0">
    ▼<generic:Series>
      ▼<generic:SeriesKey>
          <generic:Value id="UNIT" value="PC_ACT"/>
          <generic:Value id="SEX" value="F"/>
          <generic:Value id="GEO" value="AT"/>
          <generic:Value id="FREQ" value="M"/>
       </generic:SeriesKey>
      ▼<generic:Obs>
          <generic:ObsDimension value="2015-08"/>
          <generic:ObsValue value="5.2"/>
       </generic:Obs>
      ▼<generic:Obs>
          <generic:ObsDimension value="2015-09"/>
          <generic:ObsValue value="5.1"/>
       </generic:Obs>
```

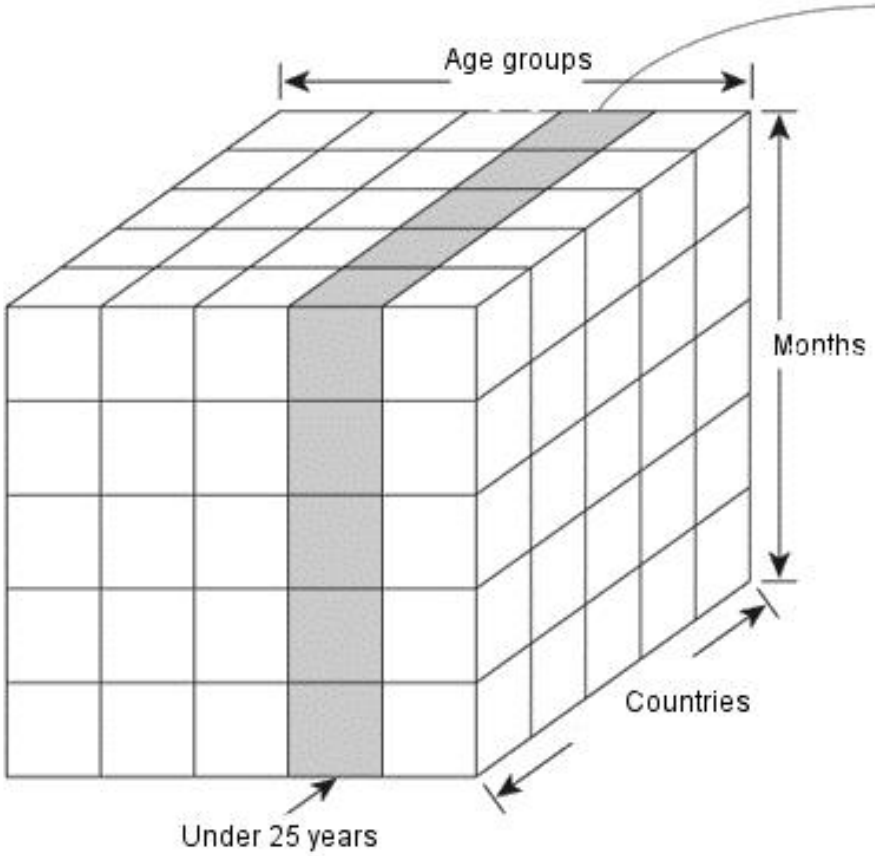
SDMX STANDARD

- But most datasets are very large and users may be seeking to download only a few series
  - This is the reason why the statistical agencies have decided to offer a genuine database service that is capable of processing specific queries

- The organisation of this database relies on a **data cube structure** commonly used for data warehousing
  - The dataset is organised along dimensions and a particular data point (stored in a cell) takes distinct values for each dimension (the combination of these values is called a key and it uniquely identities this cell)
    - Even though it is called a 'cube', it is actually multi-dimensional (i.e. allows more than three dimensions)

Slicing a data cube

- Unemployment rate of young adults (under 25 years)



	France	Belgium
2016-07	10.3	8.3
2016-06	10.1	8.4
2016-05	9.9	8.4
2016-04	9.9	8.3
2016-03	9.9	8.1

SDMX STANDARD

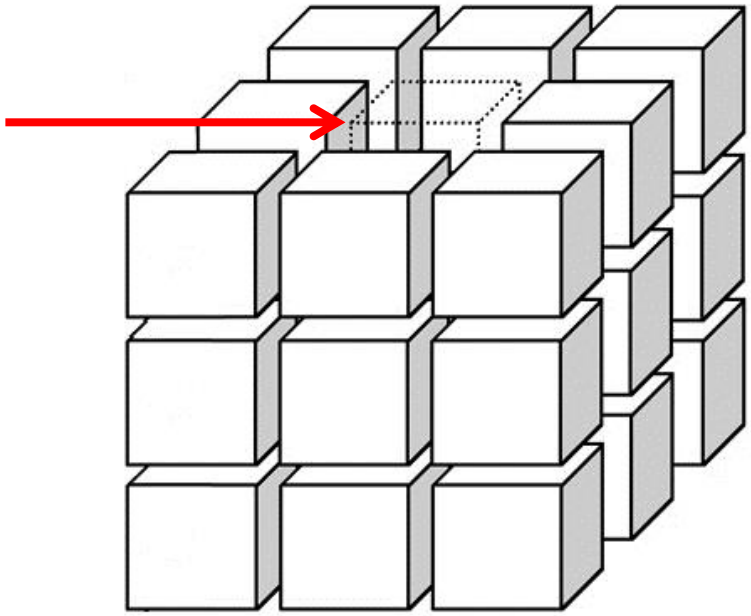
- The total number of cells of the cube in the example above is 1008
  - corresponding to all possible crossings of the variables
    - 3 age groups * 28 countries * 12 months
      - But new dimensions could be added, e.g. distinction between male and female workers or seasonal adjustment of the data

- The user should therefore identify the dimensions to be able to make a specific query
  - This is the reason why the SDMX standard provides structural metadata describing the organisation of a dataset in the form of a Data Structure Definition (DSD) file
    - giving information about the number of dimensions of the data cube, the order of the dimensions, as well as the values for each dimension

SDMX STANDARD

- The DSD gives the user enough detail to write a query for data, but it does not make any guarantees about the presence of data
  - It is quite possible that the dataset is a sparse cube (i.e. there may not be data for every possible key permutation)

```
. sdmxuse data IMF, dataset(PGI) dimensions(A1.AIPMA...) clear
The query did not match any time series - check again the dimensions' values or
download the full dataset
```



SDMX STANDARD

- The dataset in SDMX-ML format is of course flat
  - Moreover, it stores a collection of observations within each cell
    - the observations are distinguished by another dimension (often time)
- Here we observe two elements:
  - <SeriesKey>
    - an identification key with a value for each dimension
  - <Obs>
    - a set of observations with a time element <ObsDimension> and a value element <ObsValue>

```
<generic:SeriesKey>
  <generic:Value id="UNIT" value="PC_ACT"/>
  <generic:Value id="SEX" value="F"/>
  <generic:Value id="GEO" value="AT"/>
  <generic:Value id="FREQ" value="M"/>
</generic:SeriesKey>
<generic:Obs>
  <generic:ObsDimension value="2015-08"/>
  <generic:ObsValue value="5.2"/>
</generic:Obs>
<generic:Obs>
  <generic:ObsDimension value="2015-09"/>
  <generic:ObsValue value="5.1"/>
</generic:Obs>
<generic:SeriesKey>
  <generic:Value id="UNIT" value="PC_ACT"/>
  <generic:Value id="SEX" value="F"/>
  <generic:Value id="GEO" value="BE"/>
  <generic:Value id="FREQ" value="M"/>
</generic:SeriesKey>
<generic:Obs>
  <generic:ObsDimension value="2015-08"/>
  <generic:ObsValue value="7.5"/>
</generic:Obs>
<generic:Obs>
```

IMPORTING DATA FROM WITHIN STATA

- The program `sdmxuse` allows to retrieve three types of resources:
  - Data flows
    - complete list of publicly available datasets with their identifiers and a description
  - Data Structure Definition
    - structural metadata describing the structure of a dataset, the order of dimensions for the query and the distinct values for each dimensions
  - Time series data

- The syntax varies accordingly
  - `sdmxuse` **`dataflow`** *`provider`*
  - `sdmxuse` **`datastructure`** *`provider`*`, dataset(`*`identifier`*`)`
  - `sdmxuse` **`data`** *`provider`*`, dataset(`*`identifier`*`)`

- 5 providers are currently available
  - European Central Bank (ECB), Eurostat (ESTAT), International Monetary Fund (IMF), Organisation for Economic Co-operation and Development (OECD) and World Bank (WB)

IMPORTING DATA FROM WITHIN STATA

- All publicly available datasets from OECD

	dataflow_id	dataflow_description
4	FIGURE2_AEO2013_V2	Figure 2: Stock of total external debt and debt service 2013
5	TABLE2_AEO2013_V2	Table 2: GDP by Sector (percentage of GDP)
6	TABLE3_AEO2013_V2	Table 3: Public Finances (percentage of GDP)
7	TABLE4_AEO2013_V2	Table 4: Current Account (percentage of GDP)
8	WEALTH	Wealth
9	TABLE_I7	Table I.7. Top statutory personal income tax rate and top marginal tax rates for employees
10	TABLE_I3	Table I.3. Sub-central personal income tax rates-progressive systems
11	EO	Economic Outlook No 99 - June 2016
12	TABLE_II3	Table II.3. Sub-central corporate income tax rates
13	REVBOL	Details of Tax Revenue - Bolivia
14	REVHON	Details of Tax Revenue - Honduras
15	MIG_EMP_EDUCATION	Employment rates by place of birth and educational attainment (25-64)
16	QASA_7II	Institutional Investors' Assets and Liabilities
17	GVC_INDICATORS	OECD Global Value Chains indicators - May 2013
18	IO_GHG_2015	Carbon Dioxide Emissions embodied in International Trade
19	TENURE_FREQ	Employment by job tenure intervals - frequency
20	TABLE_I5	Table I.5. Average personal income tax and social security contribution rates on gross labour income
21	TALIS_EDUGPS	TALIS Indicators
22	EBDAG	Expenditure by disease, age and gender under the System of Health Accounts (SHA) Framework
23	REVNIC	Details of Tax Revenue - Nicaragua
24	ENV_KEI	Environmental - Indicators
25	TABLE_I4	Table I.4. Marginal personal income tax and social security contribution rates on gross labour income
26	BLI2013	Better Life Index - Edition 2013

```
. sdmxuse dataflow OECD, clear
```

IMPORTING DATA FROM WITHIN STATA

Data Structure Definition of the EO dataset

- The command also returns the message:

```
Order of dimensions: (LOCATION.VARIABLE.FREQUENCY)
```

	concept	position	code	code_lbl
1	LOCATION	1	AUS	Australia
2	LOCATION	1	AUT	Austria
3	LOCATION	1	BEL	Belgium
4	LOCATION	1	BRA	Brazil
5	VARIABLE	2	CGAA	Government final consumption expenditure, value, appropriation account
6	VARIABLE	2	CGV	Government final consumption expenditure, volume
7	VARIABLE	2	CGV_ANNPCT	Government final consumption expenditure growth
8	VARIABLE	2	CLF	Employment coefficient, supply
9	VARIABLE	2	CP	Private final consumption expenditure, value, GDP expenditure approach
10	FREQUENCY	3	A	Annual
11	FREQUENCY	3	Q	Quarterly

```
. sdmxuse datastructure OECD, clear dataset(EO)
```

Importing data from within Stata

- But the last OECD Economic Outlook represents more than 10000 series and about a million observations (processing time is less than two minutes though)
  - The option [, dimensions()] will "slice" the data cube to obtain only the series we want

	location	variable	frequency	time	value
1	DEU	GDPV_ANNPCT	A	1993	-.97578922
2	DEU	GDPV_ANNPCT	A	1994	2.5231408
3	DEU	GDPV_ANNPCT	A	1995	1.8182394
4	DEU	GDPV_ANNPCT	A	1996	.85380311
5	DEU	GDPV_ANNPCT	A	1997	1.9047876
6	FRA	GDPV_ANNPCT	A	1993	-.61265283
7	FRA	GDPV_ANNPCT	A	1994	2.3453856
8	FRA	GDPV_ANNPCT	A	1995	2.0850845
9	FRA	GDPV_ANNPCT	A	1996	1.388004
10	FRA	GDPV_ANNPCT	A	1997	2.3373334

```
. sdmxuse data OECD, clear dataset(EO) dimensions(FRA+DEU.GDPV_ANNPCT.)
```

Importing data from within Stata

- More options are available
  - Attributes
    - [, attributes]
      - downloads attributes that give additional information about the series or the observations, but do not affect the dataset structure itself (e.g. observations' flags)
  - Filtering the time dimension
    - [, start()] or [, end()]
      - defines the start/end period by specifying the exact value (e.g. 2010-01) or just the year (e.g. 2010)
  - Reshaping the dataset
    - [, timeseries]
      - reshapes the dataset so that each series is stored in a single variable - variables' names are made of the values of the series for each dimension
    - [, panel(*panelvar*)]
      - reshapes the dataset into a panel - *panelvar* must be specified, it will often be the geographical dimension

Concluding remarks

Remarks

- Many thanks to Robert Picard & Nicholas J. Cox for their program "moss"
- I believe that SDMX is an initiative that is worth investing in because it is sponsored by leading statistical agencies
- Some initiatives have already been implemented to facilitate the use of SDMX data for external users but they all rely on the Java programming language
- `sdmxuse` could become an alternative to private data providers (e.g. Thomson Reuters Datastream, Macrobond)

Way forward

- It might be useful to have a dialogue box with a tree structure to navigate the DSD and build queries
- SDMX standard is very likely to evolve in the coming years and more statistical organisations should join
- Not a one-man job, which is the reason why I tried to keep the ado as simple as possible, hoping more people would join the effort

REFERENCES

Resources on SDMX standard

- Official website: https://sdmx.org/
- Eurostat tutorial: https://webgate.ec.europa.eu/fpfis/mwikis/sdmx/index.php/

References on Nowcasting

- Angelini, E., G. Camba-Mendez, D. Giannone, L. Reichlin, and G. Rünstler. 2011. Short-term forecasts of euro area GDP growth. Econometrics Journal 14: 25–44.
- Barhoumi, K., S. Benk, R. Cristadoro, A. Den Reijer, A. Jakaitiene, P. Jelonek, A. Rua, G. Rünstler, K. Ruth, and C. Van Nieuwenhuyze. 2008. Short-term forecasting of GDP using large monthly dataset. ECB occasional paper series, N°84.
- Carriero, A., T. Clark, and M. Marcellino. 2012. Real-time nowcasting with a Bayesian mixed frequency model with stochastic volatility. Federal Reserve Bank of Cleveland Working Paper, N°1227.
- Stock, J. H., and M. W. Watson. 2002. Forecasting using principal components from a large number of predictors. Journal of the American Statistical Association 97: 1167–1179.