Detailed description of the data and the programs needed to replicate the figures and tables in “**Patent-Based News Shocks**” by Danilo Cascaldi-Garcia and Marija Vukotić

April 2020

Software used: Matlab R2018a, Rstudio 1.2.5019 and STATA 15.0

**Description of the original data from which the final dataset is derived:**

* The output measure is the log of real output in the nonfarm business sector (BLS: PRS85006043). The series is recovered from the Bureau of Labor Statistics (BLS).
* The hours series is the log of the total hours worked in the same sector (BLS: PRS85006033). The series is recovered from the Bureau of Labor Statistics (BLS).
* The consumption measure is the log of real personal consumption expenditures on nondurables and services (Bureau of Economic Analysis (BEA) Table 1.1.3., sum of lines 5 and 6).
* The investment series is the sum of real gross private domestic investment (BEA Table 1.1.3., line 7) and personal consumption expenditures on durables (BEA Table 1.1.3., line 4).
* The stock price measure is the log of the Standard and Poor’s 500 Composite Stock Price Index, recovered from Robert Shiller’s website.

We transform all these series into per capita values by dividing them by the BLS series of the civilian noninstitutional population over 16 (LNU00000000Q).

* The TFP measure is the log of the utilization-adjusted measure provided by Fernald (2012).
* The inflation measure is the percentage change in the CPI for all urban consumers (CPIAUCSL, St. Louis FRED).
* The federal funds rate series is the effective federal funds rate from the Board of Governors (FEDFUNDS, St. Louis FRED).
* The consumer confidence measure is taken from the Michigan Survey of Consumers as in Barsky and Sims (2011). This series is available from 1961:Q1 and, therefore, dictates the beginning period of our sample.

Industry Analysis: The CRSP data used to connect permno numbers with industry codes was downloaded via Wharton Research Data Services (WRDS). Center for Research in Security Prices, CRSP 1925 US Stock Database, Wharton Research Data Services, http://www.whartonwrds.com/datasets/crsp/. (Read more detailed explanation below)

Credit Conditions Indicators: Finally, as one of the robustness checks of our results, we control for credit conditions by considering three indicators (GZ spread, EBP, and BAA-AAA credit spread). All three credit conditions measures were taken from an online American Economic Review database provided by Gilchrist and Zakrajsek (2012), namely the file entitled GZ quarterly.csv. The start period is later than in the benchmark analysis due to data availability. In particular, the sample starts in 1973:Q3.

External Shocks: The economic shocks that we use to validate exogeneity assumption in the proxy VAR are downloaded from the Caldara and Kamps (2017) database. The measure for news about tax shocks is the proxy calculated by Leeper et al. (2013). News about government defense spending is calculated as the nominal present value of the Ramey (2011) defense news variable divided by the nominal GDP of the previous quarter, as calculated by Caldara and Kamps (2017). Oil price shocks are the net oil in- crease (three years) calculated by Caldara and Kamps (2017) based on Hamilton (2003). Monetary policy shocks are the quarterly sum of the monthly Romer and Romer (2004) variable extended by Barakchian and Crowe (2013). Tax shocks are the Mertens and Ravn (2011) unanticipated tax series. These series are in the file: Other\_Structual\_Shocks.xlsx in the DATA folder.

**Data dictionary :**

1. lvxo - log of vxo

2. y1 - real log per capita output

3. lcnds - real log per capita consumption non durables and services

4. linv - real log per capita investment

5. lh - real log per capita hours

6. GDPdefl - log GDP deflator

7. ltfp\_ua - log Fernald's TFP

8. ics - log consumer confidence (Michigan Survey)

9. infla - inflation rate

10. ffr - federal Funds rate

11. lsp\_pc - real log per capita stock price index

12. lxiq\_pc – real log per capita patent-based innovation index

13. ltfpf – log Fernald's unadjusted TFP

14. lxiq\_industry\_pc(1) - Patent-based index: finance

15. lxiq\_industry\_pc(2) - Patent-based index: manufacturing

16. lxiq\_industry\_pc(3) - Patent-based index: mining

17. lxiq\_industry\_pc(4) - Patent-based index: services

18. lxiq\_industry\_pc(5) - Patent-based index: transportation

19. lxiq\_industry\_pc(6) - Patent-based index: wholesale

20. GZ\_spread(1) - Credit Conditions: GZ Credit spread (Gilchrist & Zakrajsek, AER 2012)

21. GZ\_spread(5) - Credit Conditions: EBP (Excess Bond Premium, Gilchrist & Zakrajsek, AER 2012)

22. GZ\_spread(2) - Credit Conditions: BAA-AAA (from Gilchrist & Zakrajsek, AER 2012 database)

23. lxiq\_manuf\_pc(samp,17) - Patent-based index: Electronic And Other Electrical Equipment And Components, Except Computer Equipment

24. lxiq\_services\_pc(samp,1) - Patent-based index: Business Services

25. News about tax shocks calculated by Leeper et al. (2013)

26. News about government defense spending calculated by Ramey (2011)

27. Oil price shocks calculated by Caldara and Kamps (2017) based on Hamilton (2003)

28. Monetary policy shocks calculated by Romer and Romer (2004) and Barakchian and Crowe (2013)

29. Tax shocks calculated by Mertens and Ravn (2011)

**Construction of the aggregate patent-based innovation index:**

To construct the quarterly index we use original dataset provided by KPSS. In particular, use patents.csv and run a small STATA routine (QuarterlyXI.do). It saves the file xiq.csv. To obtain the final patent-based innovation index run Matlab file construct\_patentIndex.m. It will save the patentindex.mat file which contains the final patent-based index measure that we use in the paper: lxiq\_pc. Note that you do not need to run these routines if you are not interested in the construction of the index per se.

**Construction of the industry patent-based innovation index:**

To match industries with permno’s found in patents.csv from KPSS:

1. Take unique Permno’s used in the analysis and save into a txt file with one permno per line (instructions from WRDS on searching via permno)

2. Go to WRDS, select CRSP data download -> annual update, Stock/Security Files -> Stock Header info  
3. Download Standard Industrial Class codes for the series:

a. In Step 1, upload permno txt file created earlier

b. In Step 2, select Standard Industrial Class Code (SIC code)

c. Submit query, download results

4. SIC codes can now be used to classify permno’s into the following industry groups (as determined from <https://www.osha.gov/pls/imis/sic_manual.html>):

Agriculture, Forestry and Fishing: 100 – 999

Mining: 1000 – 1499

Construction: 1500 – 1799

Manufacturing: 2000 – 3999

Transportation, Communications, Electric, Gas and Sanitary Service: 4000-4999

Wholesale Trade: 5000 – 5199

Retail Trade: 5200 – 5999

Finance, Insurance and Real Estate: 6000 – 6799

Services: 7000 – 8999

Public Administration: 9100 – 9729

Non-classifiable: 9900 – 9999

**Construction of Firm-level Figures 2, 3 and 4 :**

**plot\_figures.Rmd**  - main file to produce figures 2,3, and 4. The file can be ran with as any Rmd file would be ran (in Rstudio) and will produce png’s of the figures. It needs 2 csv’s to run: 1) patents.csv, downloaded from KPSS 2) permno\_name\_match.csv (below). NOTE: If you do not have permno\_name\_match.csv, comment out lines 41-47 to run the file. Figures will be produced without company names.

**Main\_CGV\_REStat.m** - this is the main file which produces all the figures (except figures 2, 3 and 4 which are produced in R) and prints all tables.

**Repl\_Figs5\_6\_and\_Tab1.m** produces Figures 5 and 6, and Table 1

**Repl\_Industry.m** produces Figure 7 and Table 2, as well as Figures C1 and Table C1 from the Appendix

**Repl\_Fig8\_and\_Tab3.m** produces Figure 8 and Table 3

**Repl\_Figs\_9\_10\_11.m** produces Figures 9, 10 and 11

**Repl\_Figs\_12\_13\_and\_Tab4.m** produces Figures 12, 13 and Table 4 and 11

**Repl\_Figs\_B2\_B3.m** produces Figures B2 and B3 from the Appendix.