Great job! This could go in multiple fascinating directions... Try to pick one of two ideas & see how far you can go. Comments & suggestions below.

The Company Registers We Keep - Detecting suspicious patterns in UK corporate ownership data

Group 3: Rasim Alam, Leonie Beyrle and Taran Raghuram

1. Main Research Question

New data on company ownership published by the government of the United Kingdom provides new insights into who really benefits from private companies. Our team seeks to understand what information is available in this dataset and whether it can unearth patterns of suspicious corporate structures and illicit behavior.

2. Background and Literature Review

Detecting corruption is difficult because the outcomes are hidden by design. In particular, it has been estimated that about 70% of large-scale corruption cases involve the use of anonymous companies. These companies obscure who owns and benefits from them by (i) incorporating in "secrecy jurisdictions", such as the British Virgin Islands, where details of shareholders are not made public and (ii) by using nominees, who are listed as shareholders but have no visible connection to the actual owner / beneficiary[1]. During the last 5 years countries around the world have committed to implementing beneficial ownership registers. These vary widely in terms of what information has to be disclosed and who has access to it, but the basic idea is that companies incorporated in the jurisdiction in question are required to disclose the identity of anyone who holds over 25% of their shares or voting rights (or otherwise exercises substantial control over them)[2]. The UK's Persons of Significant Control (***PSC***) register is one of the oldest and most complete such registers and one of the few that is freely accessible to the public[3].

So far, the application of machine learning techniques to corruption research has focused on predicting corruption risks using supervised machine learning. One of the first such studies used neural networks to build an early-warning system that predicts corruption in Spanish provinces based on economic and political information about the provinces[4]. One of the most recent - and most sophisticated - studies sought to predict corruption perceptions across countries using survey data, finding Random Forest models to be outperforming SVM, which in turn performed better than neural networks[5]. However, there has been no academic research using machine learning - that we are aware of - using the UK's PSC register.

The PSC data has been used mostly for qualitative research into the ownership structures of individual companies by investigative journalists and exploratory analysis by NGOs aimed at identifying gaps in the data. The most comprehensive such analysis was carried out by anti-

[1] World Bank, Stolen Asset Recovery Initiative. 2011 "The Puppet Masters – How the Corrupt Use Legal Structures to Hide Stolen Assets and What to Do About It".

[2] Inter-American Development Bank and OECD, March 2019, "A Beneficial Ownership Implementation Toolkit", pp. 14-15, available at: https://www.oecd.org/tax/transparency/beneficial-ownership-toolkit.pdf.

[3] Global Witness, March 2020, "Patchy Progress in Setting up Beneficial Ownership Registers in the EU", available at: https://www.globalwitness.org/en/campaigns/corruption-and-money-laundering/anonymous-company-owners/5amld-patchy-progress/.

[4] Lopez-Iturriaga, F.J. and Sanz, I.P. (2017). "Predicting public corruption with neural networks: An analysis of Spanish provinces". *Social Indicators Research.* 140, pp. 1-24.

[5] Salles Melo Lima, M. and Delen, D. (2020). "Predicting and explaining corruption across countries: A machine learning approach". *Government Information Quarterly.* 37(1).

OK so the PSC data hasn't been exploited before... that's great, but you may also want to highlight why these repositories are theoretically interesting. Don't undersell your project.

corruption NGO Global Witness and volunteer data scientists at DataKind UK in 2018. They (i) gathered descriptive statistics about the dataset, (ii) developed a basic "red flagging system" for suspicious companies (where the predictors were inputs derived from guidance on corruption risks issued by e.g. international organizations, as opposed to outputs of a machine learning model[6]) and (iii) visualized company networks (using a tool they built using Linkurious, a commercial graph visualization platform)[7]. Our analysis seeks to answer similar questions to the Global Witness/ DataKind analysis using machine learning methods.

Any chance you can get a measure of company size? Or when it was founded?

3. Data Collection and Basic Descriptive Analysis

In theory, the PSC register contains a record for each beneficial owner of every privately-owned company registered in the United Kingdom. However, the PSC register dataset does not contain company names and any other basic information about the company apart from the company number, so we were required to merge this with another data set called the "Free Company Product". The free company product contains information on the registered location in a standardized format as well as the main industry the company belongs to. Together, these could constitute a helpful set of indicators for predicting outcomes or visualizing the dataset.

Great

While the PSC data looks promising, there is no official validation procedure for data entry. Companies House, the UK's registrar of companies, checks that documents submitted are complete and signed, but it does not have the capacity (or statutory authority) to verify whether the information is accurate[8]. Instead, it encourages the public to report suspected incorrect information[9] and legally requires certain "obliged entities", such as auditors and estate agents, to report discrepancies between the register and information they hold[10]. Therefore, we first need to identify how much of the dataset is usable and for what types of questions.

Our the dataset contains 5,649,275 companies with a total of 7,752,670 PSCs[11].Our initial analysis finds that:

- Over 560,000 companies declared that they have no PSC (none owns more than 25 percent of their shares) or they failed to file an owner.
- There are more than 1.9 million PSCs which are companies rather than individuals.
- Several hundred companies are registered outside of the UK, so it is unclear why they are in the UK dataset in the first place (exact number pending further data cleaning)
- About 5.3 million PSCs are individuals from England. Most of the other PSCs are from Scotland, Wales, Northern Ireland, China, and the Philippines, in that order.

Maybe they were in the UK previously?

6 Global Witness. (2018). "The Companies We Keep", pp. 20-21, available at: https://www.globalwitness.org/en/campaigns/corruption-and-money-laundering/anonymous-company-owners/companies-we-keep/.

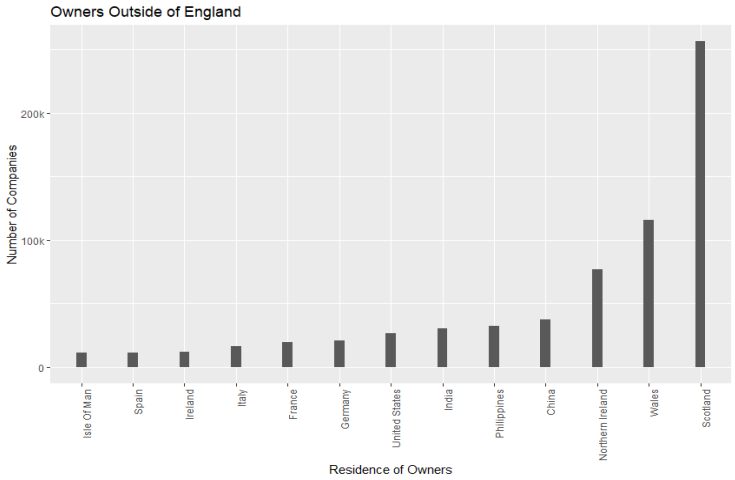
7 Ibid. p.22. See also: https://linkurio.us/

8 See Companies House disclaimer: http://resources.companieshouse.gov.uk/serviceInformation.shtml#compInfo.

9 See "Is there anything wrong on this page?" at https://beta.companieshouse.gov.uk/company/10967849/persons-with-significant-control.

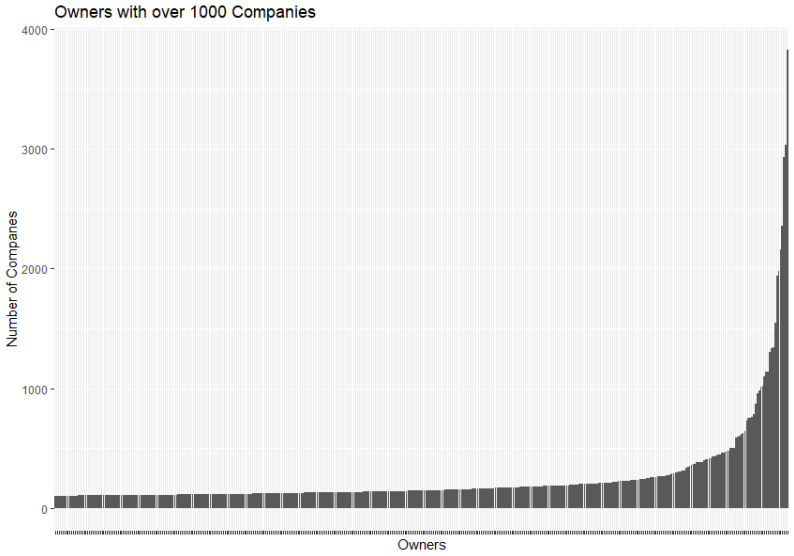
10 See guidance at: https://www.gov.uk/guidance/report-a-discrepancy-about-a-beneficial-owner-on-the-psc-register-by-an-obliged-entity.

11 As of April 20, 2020

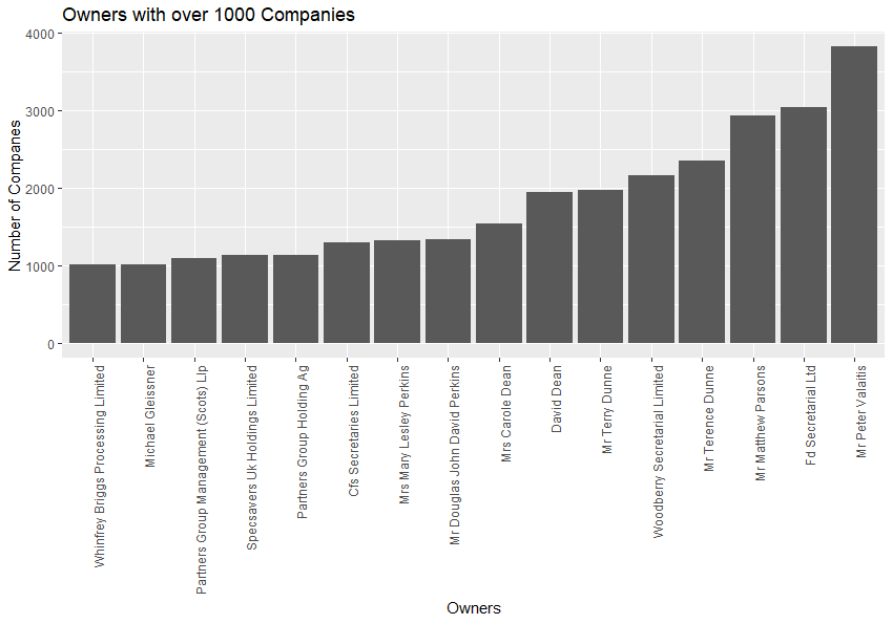


4. A closer look at the most prolific PSCs

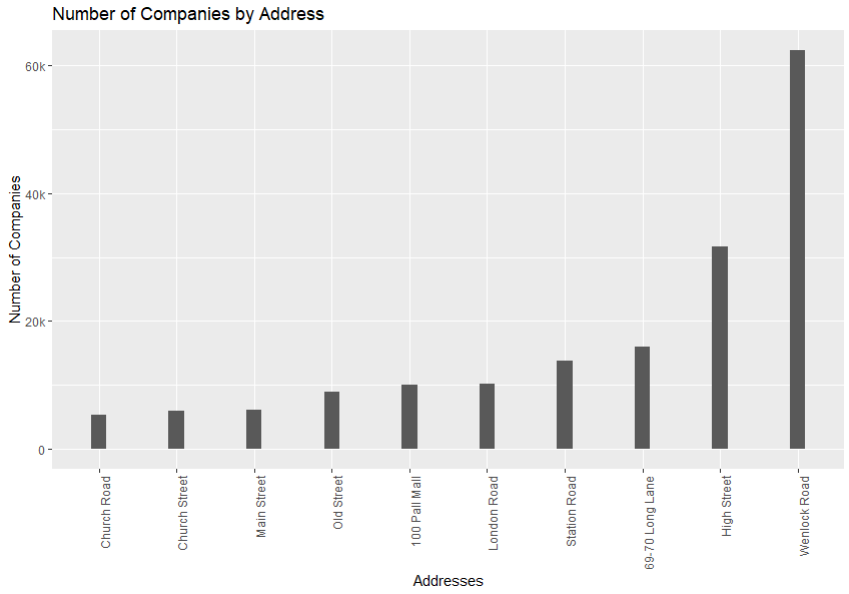
Around 104,000 companies are owned by PSCs that own over 100 companies each. As indicated by the skew of the data below, a small number of individuals or corporations own a large number of companies.



Zooming into the “high-rollers,” it appears that both companies and individuals are among them.



UK legislation only allows companies to be registered as PSCs in a limited number of very specific circumstances, raising questions about whether all of these corporate PSCs are complying with the legislation. The graph above also illustrates the register's issues with data quality, as "Terry Dunne" and "Terence Dunne" are likely to be the same person. Furthermore, the PSC register requires that companies disclose their addresses. Because the address section is a test box, many companies have mistakenly written down phone numbers, type of control, and other information. After removing those entries, we find that many companies have identical addresses listed:



Pending further investigation, this might indicate that the companies are shell companies[12] or shelf companies[13]. These have legal and legitimate uses, but are also often used to hide illicit activity.

5. Plans for improvement

We will continue to clean the data to deal with the various issues that arise because the PSC register simply lets companies write information into text boxes without verifying the information. In particular, we will need to (i) ensure we exclude all information entered into the wrong boxes, (ii) deal with typos, (iii) take into account the fact that often the same information was entered in various different ways (e.g. different versions of a name) and (iv) reconcile differences across data sets we are using (eg: only 77% of companies in PSC file had name and industry data in the free company product file, but all had that information available online)

Substantively, we will look at the following:

1. Identify ownership "dead ends": We have come across companies that are owned by a chain of other companies that eventually just ends, without any individual being named as a PSC. This can be in line with the regulations in certain circumstances, but is unlikely to be legal if the chain of ownership ends with a company incorporated in a secrecy jurisdiction.
    a. How many such chains of companies are there and where?
    b. Can our combined dataset be used to predict "red flags" (using supervised machine learning) that indicate that a company/ individual might be involved in such a "dead end" structure?
2. Create a more detailed map of where PSCs live / are incorporated using specific addresses provided.
3. Pick a number of particularly interesting/ suspicious individuals or companies and map/ analyze their ownership network more fully. This could be companies that have recently been awarded major government contracts.
4. Explore the possibility of merging our dataset with additional datasets indicating that an individual might be involved in illicit activity (e.g. disqualified directors lists, sanctions lists or Interpol red notices).

These two seem particularly promising to me...
You could also try to come up w/ a measure of "ownership structure complexity" & see how it correlates with firm characteristics.
If you do have data on govt contracts, it'd be fascinating to see if receiving a contract changes ownership structures, and if there are spillover effects.

[12] Shell companies are used as a vehicle for business transactions (e.g. to channel funds) but have no significant assets or operations themselves.
[13] Shelf companies are companies that have been created but left with no activity, i.e. "put on the shelf" to be sold later.