The Opportunities for Small to Medium Enterprises from Official Statistics

Shirley Y. Coleman
Industrial Statistics Research Unit, Newcastle University, UK
shirley.coleman@newcastle.ac.uk

Abstract

The wealth of official statistics has always been a source of fascination to academic statisticians but few of them refer to such data in their lectures. Similarly, few companies overcome the overhead investment in time and effort required to benefit from the wealth of information lying within the masses of tables available. However, there is a growing appetite for big and small data of all types and there are some excellent examples of enterprising companies creating added value from official statistics, for example www.zoopla.co.uk which offers a fascinating insight into house prices and valuations in the UK based on registered house sales and www.innovantage.co.uk which searches the web for job adverts and collates and digests them, selling the information on to recruitment agencies and others interested in labour market trends. Official statistics provide a comparison to data gathered from different sources and help identify short-comings in either direction. The Open Data Institute in London is a hub for entrepreneurs with ideas for using data. Data hacks in which data owners, designers and computer enthusiasts come together to share their expertise are gathering interest. The Royal Statistical Society recently hosted a seminar bringing together these ideas. Ongoing collaboration is a vital component of statistics maintaining its importance as a subject and of statisticians continuing to play a major role in the data revolution.

Key Words: entrepreneurs, official statistics, open-data, SME

1. Introduction

In spite of the vast importance of business statistics for policy making at all levels of society, this branch of statistics has received comparatively little research and methodological attention. The wealth of official statistics has always been a source of fascination to academic statisticians but few of them refer to such data in their lectures. Similarly, few companies overcome the overhead investment in time and effort required to benefit from the wealth of information lying within the masses of tables available. However, there is a growing appetite for big and small data of all types and there are some excellent examples of enterprising companies creating added value from official statistics, for example www.zoopla.co.uk which offers a fascinating insight into house prices and valuations in the UK based on registered house sales.

One of the reasons for the lack of use of official business statistics is that data owners, data users and entrepreneurs tend to be isolated in their own silos and do not often meet. Yet there have been some excellent examples of mutual benefit when these different groupings work together and it is evidently important to bring people together to discuss their ideas. Similarly, it is essential to advertise the success of partnerships to inspire others and suitable market places for such exchanges are beginning to occur. Another major reason for lack of use of official statistics by business is poor numeracy and statistical competence within the business community.

This paper reports some of the ways in which the barriers to better business use of official statistics are being tackled so that small and medium enterprises (SMEs) in particular can realize the many opportunities available to entrepreneurs. Section 2

introduces two networks that aim to bring people together and outlines recent seminar presentations to show the breadth and depth of potential and actual projects. An example of an SME making a successful business venture using open data and official statistics is given in section 3 along with another example, the *data hack*, which is an intriguing way to bring together diverse communities to uncover information buried in data. The numeracy shortcoming in the UK is being addressed in a major initiative called *getstats* by the Royal Statistical Society (RSS) and section 4 looks at how statistics, open and big data are starting to work together to explore their synergies. Inspirational programmes such as the worldwide movement towards drawing young people into computer coding and data mining will help in the future to overcome the barriers preventing fluent use of official business statistics and the final section considers how young people are becoming involved and offers some conclusions.

2. Business Statistics Networks

The links between the Office for National Statistics and the RSS have always been valued and nurtured. A major project resulted in an interactive website for all users of official statistics, called *statsusernet* which has just celebrated its first anniversary. This provides a forum for communities interested in official statistics. For example the Business and Trade Statistics community aims to promote dialogue, share information and maintain close liaison between the producers and users of official business and trade statistics.

Since its initiation in 2008, the European Network for Better Establishment Statistics (ENBES) has worked on advancing exchange between practitioners, methodologists, and academics on matters relating to business statistics. In parallel with - amongst others - the European Commission's MEETS programme and the FP7 research project BLUE-ETS that were initiated at approximately the same time, ENBES is endeavouring to bring business statistics closer to its users, as well as helping understand user needs for business statistics better.

The RSS Official Statistics Section (OSS) is concerned with the collection, analysis and interpretation of official statistics of all kinds including social, economic and environmental statistics. The RSS Quality Improvement Section (QIS) was originally a spin-off from the Business and Industrial Section motivated by the increasing awareness of the important role that statisticians should play in quality improvement. The QIS aims to spread awareness of the powerful contribution that understanding variation and statistical methods makes to quality improvement and good management. Its objectives are to help members contribute effectively to the measurement and improvement of quality and performance.

The OSS is interested in issues around collecting high quality data and presenting it clearly and accurately; the QIS is interested in supporting statisticians using data to improve the effectiveness and profitability of enterprises in all sectors. The seminar "Statistics: About Businesses, for Businesses" was co-organised by ENBES and the OSS and QIS as an example of bringing together different interest groups. This was the first time that the two sections have held a joint meeting and reflects the growing interest in business use of data and statistics. The seminar was held at the RSS in London on February 19, 2013, and offered a balance of up-to-date presentations including an overview of the field, case studies of data use and issues arising therein, importance and potential of new data sources, as well as a summarising keynote talk.

A description of parts of the seminar will be instrumental in showing the range of topics and issues involved in collecting and producing official statistics on businesses and, on the other hand, to showcase potential uses of official statistics and businesses own data for entrepreneurs, for research and for policy making. The seminar brought

together practitioners, users and methodologists to improve mutual understanding and stimulate more effective use of official business statistics especially by SMEs.

Gordon Blunt (Gordon Blunt Analytics Ltd) gave an interesting presentation on the way that data already held by businesses could be used in their management. With the power of modern hardware and software, many companies nowadays possess large databases. Given that there is a computer on virtually every desktop, it should be easier than ever for companies to make good use of their data. However, there are many pitfalls on the road to insight, such as the quality of the data, the use of operational systems that were never designed for analysis, legacy systems that do not fit well with modern software, missing or incomplete records and so on and so on. One of the most common pieces of business software is the spreadsheet. Spreadsheets were first introduced in the 1970s, and, in their early days, had limited data handling capabilities. Given the ubiquity of modern desktop computers and the powerful spreadsheet software now available, anyone who has a data set can produce their own analysis. This 'democratisation of analysis' can give statisticians many opportunities, but it may also lead to problems. Making sense of data has always needed appropriate skills and knowledge, regardless of the size of the data set. Further, statisticians have always known the importance of examining their data, but there are particular problems with large data sets, and visualisation is critical – both for examining structures and identifying data problems. Errors in a large data set may not be trivial to correct; indeed, the larger the data set, the harder the task. The talk suggested ways to ensure that companies can make the best use of data. Use of external data to help develop insight may be the only way a company knows how it is performing compared to its competitors. In summary, he argued that data cleaning is by far the biggest part of a statistical job, and that getting the right visualisation can be one of the most helpful things for a client.

Stuart Coleman (Open Data Institute) introduced his institute (ODI) which was founded in 2012 by Sir Tim Berners-Lee and Professor Nigel Shadbolt as an independent, non-profit, non-partisan company limited by guarantee. Its aims include unlocking the supply of data, including data from the private sector, and communicating its value to potential users. Synergy between the ODI and RSS is discussed in section 4.

Stephen Penneck (President of the International Association for Official Statistics and former Director General of the Office for National Statistics) looked at the evolution of statistics on businesses from trade protection in the 19th and early 20th centuries to an expansion of detailed product information for use by business in the 1970s, and then an increasing focus on macroeconomic outputs. He highlighted the opportunities raised by Open Data initiatives, and challenged the audience with questions on whether government is missing anything with the current macroeconomic system, and whether availability and use of business data are joined up across Europe.

The day was very useful in raising the profile of business statistics, showing their range and generating ideas on how they could be better communicated and utilised. The presentations are available via ENBES at [2] and on StatsUserNet [3].

3. Case studies

More use of official statistics needs to be encouraged and a good way of doing this is to present case studies where SMEs have added value to official statistics and been able to build a business based on selling on the augmented information. Two case studies are considered in this section. It is hoped that more will become available and will be instrumental in starting an avalanche of fruitful applications in this area.

3.1 Job Vacancies

Innovantage (see www.innovantage.co.uk) is an SME which searches the web for job adverts and collates and digests them, selling the information on to recruitment agencies and others interested in labour market trends. They collate nearly all on-line job adverts using a proprietary web search system. This results in a database consisting of approximately 1.5 million job adverts from nearly 200 job boards every month which makes it an extremely rich source of intelligence. A major barrier to providing high quality is that a single job advert can be posted on multiple job boards, by multiple recruitment agencies and reposted multiple times by job boards to drive up their apparent traffic. Stripping out all of this duplication in order to arrive at the true number of job vacancies is a significant statistical exercise [4]. In the project, job vacancy information derived from on-line job advertisements were compared with official estimates of job vacancies supplied by the Office for National Statistics using their Vacancy Survey and Labour Force survey. There were encouraging similarities between them, but the comparison revealed considerable challenges in de-duplicating and classifying vacancies information derived from internet job sites. The official statistics data was used to develop new methods of de-duplication, identify unexpected data quality issues and to inspire a radical way of overcoming a barrier to labour market intelligence that was thought to be insurmountable by all participants in the industry, namely the issue of commercial relevance and timeliness. This interesting case study was presented at the seminar discussed in section 2.

3.2 Hacks

Hacks are events in which data suppliers, designers and analysts come together to share their expertise and create added value. *Rewired State* designs and creates hack events that bring creative developers and industry experts together to solve real-world problems [5]. It is fascinating how little statistical input there is at present in these events. Statisticians need to reclaim their place in this new movement to make best use of data. For example, at a Culture Code Hack held in Newcastle experts presented data from lending libraries, songs from different parts of Northumberland and location and demographics of theatre goers, amongst other data. The author was the only statistician at the event and presented data from the ENBIS challenge [6]. Computer scientist hackers then spent the next 24 hours finding innovative ways to illustrate and interpret the data producing interactive maps showing who was looking at computer software library books, where and for how long; who went to which theatre performances and what melodies to expect in different Northumberland villages.

4. Synergy between statistics and the open data movement

4.1 Synergies between institutions

The Open Data Institute's mission [6] is to catalyse the evolution of open data culture to create economic, environmental, and social value. It will unlock supply, generate demand, create and disseminate knowledge to address local and global issues. It will convene world-class experts to collaborate, incubate, nurture and mentor new ideas, and promote innovation. It will enable anyone to learn and engage with open data, and empower their teams to help others through professional coaching and mentoring.

One of the first ODI events was a *Midata Hackathon* to explore the future of personal data. Developers, designers and data experts explored a future in which consumers will have easy access to information collected about them by businesses. One of the example questions posed was about Health and Fitness: How could you use personal data to help people lead healthier lives? What can you work out from exercise, food purchases, and other financial information?

The *getstats* initiative [7] is a 10-year campaign mounted by the RSS with the support of the Nuffield Foundation to improve how we handle numbers – the practical

numbers of daily life, business and policy. Statistics are tools that turn data into useful information. They give numbers meaning and help to decode complexity.

There is clearly considerable synergy between the ODI and RSS in their missions to capitalize on data and the perceptive analysis of data. Discussions on ways in which statisticians can help people relish the collection and interpretation of data and use analytical methods to gain insight from the data are ongoing and are expected to be very fruitful.

4.2 Synergy in research

Newcastle University have submitted a bid to the research councils to create a Centre of Doctoral Training (CDT) in "Cloud Computing for Big Data". If funded, this would allow 10 students per year, for 5 years, to begin a PhD in this area. The motivation is that while cloud computing offers the ability to acquire vast, scalable computing resources on-demand, its full potential for extracting knowledge from data has not been realized outside a few large organisations; many organisations fail to realize the potential to be transformed through extracting more value from the data available to them. The CDT aims to overcome this barrier by combining three approaches:

a) Research to develop a deep level of understanding of the theory and practice of big data analysis, creating innovative solutions where none exist.
b) Collaboration between computer scientists, mathematicians and statisticians to develop methods to understand and model data so as to reduce the amount of computation required, and to gain deeper insights into that data. The CDT will create researchers and practitioners who can bridge between the design of scalable algorithms and the underlying theory in the modelling and analysis of data.
c) Producing a stream of high-quality graduates with multi-disciplinary expertise in the mathematics, statistics and computing science of extracting knowledge from big data, and practical experience in exploiting this to solve problems across a range of application domains. Currently industry is suffering from a real shortage of expertise in this area.

A combination of statistical and computing skills are required by researchers in this area. "Statistics for big data" involves numerical linear algebra, likelihood, regression, multivariate data analysis, graphs and graphical models, exchangeability, sufficiency, hierarchical modelling and Bayesian inference. "Big data analytics" involves distributed computation of statistics, model fitting and analysis, and visualization techniques. "Time series data" requires a knowledge of time series models and analysis, on-line versus batch learning, state space models, Kalman filter, particle filters, trend analysis and approximate counting. "Programming for big data" includes R, Java, data mining, algorithms, databases and query languages, nosql and graph databases. "Cloud computing" involves distributed computing, cloud architectures, distributed algorithms, virtual machines and scalable computing patterns. In addition, researchers will take modules on Entrepreneurship and Innovation.

5. Conclusions and future work

Statisticians need to claim their place in the data revolution. A route to this is in ensuring young statisticians are inspired to work within the area. An example of how this enthusiasm can be nurtured is the Young Rewired State [8] initiative just starting in UK schools. In 2009, Rewired State [5] decided to host an event called "Young Rewired State", a weekend hosted by Google in their London offices intended to introduce open government data to the coding youth of the UK. Their website notes:

"With great excitement and anticipation of meeting these young programmers we flung open the doors with a limited capacity of 50, due to the restrictions at Google

London offices.

Three young people signed up.

As we called schools and scoured the internet we realised that there was a far bigger problem than young people not engaging with open government data. That was the lack of young programmers in the country, and the fact that we were still left with isolated kids, teaching themselves how to code in their bedrooms – terrifying their parents that they were up to no good. Schools, would often identify a lone individual who might be interested – but beyond that they could not help as they had long since stopped teaching programming.

We then spent three months focused on finding the founding fifty, and with huge relief and even more anticipation, we brought them together. We ran a weekend, with mentors and government data experts on hand to help, and watched as they collaborated and created a blistering array of apps and websites, all using open government data.

In front of our eyes a community was born, something that was so needed – never again would these young geniuses be coding alone, from now on they had their peers and mentors to be a part of their education and maturation into engaged civic programmers."

Networks such as ENBES and facilities such as *statsusernet* are important mechanisms for aiding business use of official statistics. Entrepreneurs are seeing the opportunities for adding value to open data and making business use of small and big data. Statisticians need to keep in contact with experts from other fields to ensure that they keep pace with the changes taking place in the way data are perceived. The RSS and universities are starting to see ways of developing statisticians for the new data era. Working alongside data owners, designers, hackers and programmers, statisticians can help statistics as a subject become firmly established as part of the data revolution.

References

[1] Royal Statistical Society http://www.rss.org.uk accessed April 2013

[2] Seminar presentations available at http://enbes.wikispaces.com/Statistics+for+businesses+about+businesses accessed April 2013

[3] Seminar presentations and much useful information available at http://www.statsusernet.org.uk/StatsUserNet/LibraryFolders?LibraryKey=c17d3aac-1b27-4b67-9077-c8ed66fbad87 accessed April 2013

[4] Further information via http://marriott-stats.com/Testimonials accessed April 2013

[5] Hacks are described at http://rewiredstate.org accessed April 2013

[6] Presentation at CultureCodeHack http://vimeo.com/39778333 accessed April 2013

[6] Open Data Institute http://www.theodi.org accessed April 2013

[7] RSS *getstats* http://www.getstats.org.uk/about/ accessed April 2013

[8] The coding revolution for young people https://youngrewiredstate.org/about-yrs accessed 1st April 2013