DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 32, No. 4, 2016, pp. 849–865, http://dx.doi.org/10.1515/JOS-2016-0044

Data-Mining Opportunities for Small and Medium Enterprises with Official Statistics in the UK

Shirley Y. Coleman[1]

There is a growing interest in data amongst small and medium enterprises (SMEs). This article looks at ways in which SMEs can combine their internal company data with open data, such as official statistics, and thereby enhance their business opportunities. Case studies are given as illustrations of the statistical and data-mining methods involved in such integrated data analytics. The article considers the barriers that prevent more SMEs from benefitting in this field and appraises some of the initiatives that are aimed at helping to overcome them. The discussion emphasizes the importance of bringing people together from the business, IT, and statistical worlds and suggests ways for statisticians to make a greater impact.

Key words: Entrepreneurs; company data; open data; analytics; impact.

1. Introduction

There is a growing trend within the business world for utilizing big and small data of all types and there are some excellent examples of enterprising companies creating added value by data mining with official statistics. Small and medium enterprises (SMEs) are relatively slow to adopt such a data-focused approach, but competitive pressures on businesses are creating awareness and interest in this area. It is timely therefore for statisticians to consider the issues involved and to help SMEs come to terms with the new ways of thinking about data.

In spite of the vast importance of official statistics for policy making at all levels of society, there is comparatively little use of such data in the business world. Large and small companies and industries in all sectors can benefit from census and business data collected by National Statistics Institutes (NSIs). Some large companies, such as Google and Amazon, have led the way in data analytics, for example using customer data to build recommendation systems that predict what the customer wants or needs next. In quality improvement initiatives such as Six Sigma (see, for example, Snee and Hoerl 2004), large organizations are earlier adopters than SMEs, and this is the case for integrated data-analysis initiatives. A study by the European Union (EU) funded BLUE-ETS project found that "small businesses use data modestly" and "the main obstacles preventing businesses from using NSI statistics (more intensively) include. . . lack of interest" (Bavdaž 2011).

This article takes a closer look at how statisticians can help SMEs use improved data analytics. Specifically the article focusses on the benefits of integrating internal company

[1] Industrial Statistics Research Unit, School of Mathematics and Statistics, Newcastle University, Newcastle upon Tyne, NE1 7RU, UK. Email: shirley.coleman@newcastle.ac.uk

data with publically available open data. Such integration provides added impetus to SMEs showing interest in improving their data capability to their business advantage.

The contextual characteristics of SMEs affect their willingness to embark on data-analysis initiatives. Statisticians need to understand these and work with SMEs to raise awareness of the opportunities. Case studies of successful applications of statistics with company and official data are a vital way to provide both motivation and a roadmap. Other mechanisms are needed, however, and some potential routes to reach out to SMEs are discussed in the present article.

Expanding the utilization of official statistics is timely from the official statistics point of view as governments become increasingly more careful with their money, looking for evidence of return on investment. If business usage is increased and widened, then this helps to justify the costs of compulsory collection of business and other data.

The objectives of the article are to:

- showcase examples of SMEs realizing business advantage from working with integrated data and highlight the appeal of such projects for statisticians;
- review the barriers specific to SMEs that dissuade them from utilizing integrated data analytics;
- explore and appraise some activities that aim to overcome the barriers;
- make recommendations for statisticians to help SMEs take advantage of the increased availability of data.

The article is structured as follows: Section 2 provides a brief background including the contextual characteristics of SMEs that influence the way they relate to data, the wider definition of official statistics used in the article, the meaning of terms such as data analytics, and the methods of statistical analysis and data mining that are particularly useful. This is followed by illustrative examples with increasing depth of analysis, ranging from descriptive analytics of official statistics to prescriptive analytics of integrated internal company data and official statistics used for calibration. Section 3 reflects on the issues raised by the examples and considers barriers to SMEs embarking on integrated data analytics. Various engagement activities such as networks, data hacks, and institutional initiatives are appraised in the final Section 4, and the article concludes by suggesting areas for increased involvement by statisticians, particularly in terms of business impact and positive benefit to society.

2. SMEs and Integrated Data Analytics in Theory and Practice

2.1. Background and Definitions

SMEs make up 99% of an estimated 19.3 million enterprises in the EU and provide around 65 million jobs representing two-thirds of all employment (World Bank 2015). The average European business provides employment for four people, including the owner or manager. SMEs are characterized by an innovative idea or product, usually created by the owner. There are unlikely to be many data-competent staff. Case studies from the EU-funded BYTE project show mixed feelings about using data, including worries about security and a wariness of investing resources and time in developing the necessary data-analytic skills (BYTE 2015b). SME business needs, however, are to reach appropriate markets and to make an impact by showing the importance of their product. Identifying areas where their specialist skills can add value to a free resource should be a major focus. It is therefore likely that if SMEs can see the benefits of data mining and using different sources of data, then they will be motivated to get started.

The term “official statistics” usually refers to data collected by governmental NSIs. The data are characterized by their high quality, their standardized definitions and thorough descriptions, and considerable harmonization between departments and countries. Official statistics include extensive demographic data from census and government department sources as well as data from compulsory business surveys. Although official statistics are generally available in most countries, this article refers mainly to the UK because the UK is particularly interested in promoting the use of official statistics. For example, the Royal Statistical Society (RSS) gave ten recommendations in its Data Manifesto (RSS 2014) for widening the use of data and official statistics in particular; the National Statistician, John Pullinger, is currently reviewing the Approved Researcher process by which access to government data are selectively allowed, with the aim of widening usage of official statistics and making access more straightforward. In addition, the UK Open Research Data Forum produced a Draft Concordat on Open Research Data (Research Councils UK 2015) which aims to ensure wherever possible that research data are made openly available for others to use.

Alongside official statistics, it is sensible also to consider the large quantities of data made available under the umbrella of open data. The Open Data Institute (ODI) in the UK was founded in 2012 by Sir Tim Berners-Lee and Professor Nigel Shadbolt as an independent, nonprofit, nonpartisan company limited by guarantee. ODI aims include unlocking the supply of data, including data from the private sector, and communicating its value to potential users; catalyzing the evolution of open data culture to create economic, environmental, and social value, generating demand, creating, and disseminating knowledge to address local and global issues (ODI 2015).

Open data includes data collected in projects receiving public funding, although clearly the published data may differ in depth, granularity, and coverage from the actual research results. Official statistics are particularly useful for SMEs because they are consistently and reliably available. However, both open and official statistics data are free and are a potentially useful resource for businesses to utilize. In the rest of the article we will distinguish between open data, including official statistics, and internal company data, which is generally closed and for company use only.

Data science is big business. It is a combination of data awareness, skills in data storage and manipulation, and development and application of analytical methods. Considerable insight can be obtained from descriptive analytics in which data are summarized, sliced and diced to produce tables applicable for business consumption. Descriptive analytics looks at historic data to find out what has happened. Predictive analytics goes a step further and uses analytical methods to extract patterns in the data and make predictions about the future. Predictive analytics involves data-mining techniques such as decision trees and cluster analysis, and statistical methods such as regression modeling and principal components. Real data are often very noisy, but predictive analytics makes it possible to compose rules for achieving targets and to detect relationships that are useful, for example in customer segmentation (Ahlemeyer-Stubbe and Coleman 2014). Prescriptive analytics not only predicts what will happen in the future but also considers what decisions to make as a result of the predictions. It also looks at the explanations for and implications of the predictions. The application of data-mining techniques and statistical methods is referred to as data analytics.

Internal company data include sales records with details of customer demand and location, quantities and values of transactions, time stamps, and feedback. It is a rich source of information even on its own, as exemplified in Ahlemeyer-Stubbe and Coleman (2014, chap. 10.3). In conjunction with official statistics they become even more powerful, as is shown in the examples below. A major issue is the integration of the datasets; they need to be matched, the data records may have different key identifiers, they may differ in temporal and spatial granulation and in completeness. Kenett and Shmueli (2015, chap. 10) address these issues and also note the strengths of the resulting integrated data. Applying data-mining techniques and statistical methods to integrated data is referred to as integrated data analytics.

‘Big data’ is an increasingly popular phrase that can apply to some SME data. Big data are characterized by having large VVV – because it may include a great *Variety* of data types, including counts, measured data, 2D and 3D spatial and image data; it may arrive with great *Velocity* and with great *Volume*. ‘Great’ is defined as requiring more than commonly available storage, manipulation, and analytical facilities. Big data has entered into everyday usage; for example, it was mentioned in a recent UK government annual budget statement. Such increased exposure has aroused curiosity and people, especially those working in SMEs, are becoming increasingly interested in hearing what it is all about. Hence now is a good time for statisticians to work with SMEs and encourage them to think more about analyzing their data.

EU funding recognizes the importance of data; for example, the European Data Forum project (European Data Forum 2015) arranges an annual event where anyone (organizations, charity, policy makers, and researchers) can discuss the challenges of analyzing big data and the development of emerging data economies. The special focus is on SMEs as they are the main players in the emerging data economies. The forum discusses the challenges of changing business models and issues of legality and privacy in big data; it helps organizations understand the risks and benefits of data analysis. Conference proceedings include a list of relevant EU-funded projects and other groups that showed an interest in the discussions (European Data Forum 2014). EU projects such as BYTE (2015a) and BIG (2015) concern data but neither of these focusses specifically on official statistics and the opportunities for SMEs.

2.2. Illustrative Examples

2.2.1. Section Overview

In this section, the aim is to show a range of applications of official statistics to help businesses. The first example looks at companies that have used descriptive analytics to good effect without having to apply any specialist statistical skills. The only requirements are that the companies are aware of the availability of the official statistics, know how to access them and are motivated to use them because they are confident of their business impact. The second example looks at a company that is just starting to develop their data-analytic capabilities. It shows how official statistics can be integrated with more specific company data and make a distinct improvement in the value of their data. It shows the need for data-manipulation skills and demonstrates why SMEs may find it difficult to embark on this sort of work. The third example looks at a company that has a specific goal in mind and has access to statistical and data-mining expertise. It shows how official statistics can be used in a calibration capacity to check and corroborate data obtained in other ways.

2.2.2. Official Statistics as Descriptors

There are great opportunities for SMEs to use official statistics in a specific way and sell on the user-friendly information. Publically available data on house sales from the UK land registry, for example, have been transformed by www.zoopla.co.uk into a fascinating insight into house prices and valuations. Using demographic data from census relating to local areas, Zoopla presents a whole range of interesting facts about any chosen housing area. The data are fundamentally freely available and the value added by the company derives from the automatic data accumulation, accessibility and presentation designed by them. The business model is to attract visitors to the website and sell advertising.

Official statistics can be used to summarize the current situation and predict the likely trajectory of market sectors. For example, consider the announcement:

> *The UK maritime industry contributes up to £14 billion pa; expected to rise significantly. Direct employees include around 260,000 people.*

Source: UK Government (2014).

These figures are highly valued by an SME engaged in the maritime equipment industry. In fact, this SME is so enthusiastic about data that they are currently engaged in a two-year Knowledge Transfer Partnership (KTP 2015) which aims: “to use statistical and data-mining techniques to facilitate the further development and enhancement of the company’s *enginei* monitoring system”.

The project includes data mining company data combined with using official statistics to develop a new product and reach out to new areas of interest to its customers. Sensors provide big data on fuel consumption, speed over ground, and geographical location, and these data are combined to detect the mode of operation of a ship automatically. An important application area is emissions control, which is becoming increasingly more highly regulated (Coleman et al. 2015b). Emissions can be calculated for specific engines by equations based on engine characteristics. However, the company needs to demonstrate that their calculations are correct. Conversion factors from international standards are used to verify the calculations. The conversion factors apply to general classes of engines, whereas the calculation relates to specific engines. Once verified, the emissions calculation can be used to give tailored responses for specific engines. In this application, the expected emissions for different modes can be calculated and used to inform operational decisions.

Many SMEs market a niche product appropriate for a subset of the population. Official statistics can clearly be used to identify and prioritize areas for business expansion. For example, an SME designed an expert system to help people who are unable to carry out one or more activities of daily living. There are a large variety of products available to help people walk or wash or whatever they need, but people do not know which product to choose. The expert system identifies suitable products that match the person's needs exactly. The company has many millions of records of searches dating back over the last ten years and is engaged in a two-year Knowledge Transfer Partnership (KTP 2015): "to exploit existing big data arising from decision making, by appropriate mining, monetization and marketing and to extend company offering from public to new markets including the private sector."

Using official statistics the company can find, for example, the percentage of people with disabilities in a particular geographical area and how the percentage is changing with time. This information gives insight into likely demand and a means to estimate their market penetration (Coleman et al. 2015a).

2.2.3. Segmentation Using Company Data and Business Statistics

Company data are usually private and closed for company use only. In conjunction with official statistics, the full breadth of demographic information collected in census and surveys can be used by the company to increase their understanding of the customer base. A project was carried out in a Knowledge Transfer Partnership (KTP 2015) with the aim: "to develop analytical skills applicable to gas flow data to improve the efficiency of gas distribution at minimum cost to the system."

Energy companies have big data from operations and system control, including measurements such as flows and leakage, and records relating to asset condition. In addition, they have copious data on customer demand. It is very important to be able to predict demand as this ensures minimum stress on supply systems, less storage requirements and more efficient and reliable energy delivery. Customer demand varies according to season, day of week, time of day, alternative energy options and prices, domestic or business/industrial use and type of customer. Demand forecasting based on seasonality and customer type is well established; however, less work has been carried out on the effect of customer characteristics on demand. Official statistics provides an opportunity to develop this area in greater detail.

As part of the KTP, company data on gas demand are plotted geographically to demonstrate their variability. There are distinct patterns, and interest focusses on how these patterns relate to demographic characteristics. A literature review suggests which characteristics are most likely to be important and relevant indicators are extracted from official statistics. Income and an index of deprivation in housing, for example, are then superimposed onto spatial maps of the demand data (Coleman and Yabsley 2015). Further official statistics are collected and prescriptive analytics are carried out. Official statistics tend to be correlated and techniques such as partial least-squares analysis can be used to find powerful predictive models for demand. By integrating customer data and publically available statistics, customers can be clustered into segments; for example, Fontdebaca et al. (2012) identified six segments of water users. Such segmentation helps the utility company to improve its planning and the reliability of its supply. Such analysis is very useful in providing the company with insights and motivating future interventions to give a better service to their customers.

This sort of analysis involves many steps and different skills. It is not surprising therefore if SMEs are reluctant to embark on such exercises. For example, in the KTP, customer data are logged in terms of postcode; postcodes need to be mapped to government-defined local area codes and data need to be extracted from official statistics for those codes. All the information is available from Office for National Statistics (ONS) websites, but the files tend to be very large and require preprocessing to be accessible on a personal computer. Mapping postcodes to local area codes requires data manipulation, for example using lookup tables in Excel. The modeling requires statistical software, either a package such as SPSS or R algorithms, the use of which has to be learned. There is also an overhead of dealing with messy data in formats that may need converting; for example, the customer-demand data were in pdf files that needed to be converted to spreadsheets before the figures could be extracted. However, the effort involved is well repaid by the interesting results.

2.2.4. Business Statistics as Calibration in Job Vacancies

Innovantage (see www.innovantage.co.uk) is an SME which searches the web for job advertisements and collates and digests them, selling the information on to recruitment agencies and others interested in labor-market trends. The SME collates nearly all online 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. This example was presented as a case study at an RSS seminar entitled “Statistics: About Businesses, for Businesses” described in the discussion below (ENBES 2013).

A major barrier to providing high-quality information is that a single job advert can be posted on multiple job boards and by multiple recruitment agencies; furthermore, it can be reposted multiple times by individual 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. In the project, job vacancy information derived from online job advertisements was compared with official estimates of job vacancies supplied by the ONS 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 vacancy information derived from Internet job sites. The official statistics data were used to develop new methods of de-duplication, identify unexpected data-quality issues, and to inspire a radical way of overcoming a barrier to labor-market intelligence that was thought to be insurmountable by all participants in the industry, namely the issue of commercial relevance and timeliness. After developing the methodology for establishing a high-quality database of job vacancies, the company was able to carry out data-mining activities, including segmentation and modeling, to clarify the underlying trends in the market, identify patterns and predict shortfalls and demand, all of great business value.

In this example, official statistics are used to calibrate a dataset derived from data from other sources. Other examples of calibration include Dalla Valle (2014), in which a nonparametric Bayesian belief network is built using data from an association that collects information about companies. Information about companies from an official statistics source is then used to see what characteristics a set of firms should have in order to perform similarly to the firms described in the official statistics. In this way the association's data can be calibrated and exceptional companies can be investigated.

Company data are most powerfully used in conjunction with official statistics, as discussed above. However, company data have usually been collected for other purposes, such as immediate operational control, and whilst the quality may be good enough for a specific task it often suffers from difficult issues such as the overall 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. Dalla Valle and Kenett (2015) consider the difficulties of calibrating official statistics with such variable datasets. In addition, they propose ways of improving the information quality (InfoQ) of the official statistics used in this context. Their analyses show important benefits for the decision makers in the companies involved. The case study of Innovantage was selected as an example, however, because it shows how an SME is building its business specifically on the data it manipulates.

3. Barriers to Adoption of Integrated Data Analytics in SMEs

3.1. Lack of IT Skills

The examples in Section 2 require a range of analytical skills and knowledge, including knowing how to access and manipulate suitable data and apply appropriate statistical and data-mining skills. Many large organizations have moved into data analytics by partnering with an IT company that can provide these capabilities. For example, the supermarket Tesco is now well known for its extensive use of customer data, but Tesco's first role as a company is as a supermarket. Realizing the potential of data mining, Tesco started working with Dunnhumby in 1995; Dunnhumby are specialists in data collection and analysis and they helped Tesco to roll out the loyalty-card program. The loyalty cards encouraged customers to shop at Tesco but also they were the means of recording the vital data that Tesco needed to explore and understand their customers, create target advertising around the items they bought, utilize clever stocking which would increase the number of sales in items and so on. Tesco eventually bought Dunnhumby and they would not have been so successful without the help of an outside organization (Mirani 2015).

Another example is Spotify. The company offers users an ad-free music streaming service on a subscription basis. Their main business is music; however, they use data-mining techniques to link songs together and they bought Echo Nest, a company involved in music analytics, to help them do this. One of the main reasons for the company having a turnover of £131.4 million in 2013, a 42% increase from the previous year, was because of their effective utilization of data analysis (Duedil 2015). In 2013, 4.5 billion hours of music were listened to; Spotify was able to offer a better experience to users, with recommendations of similar music in addition to the music they are currently listening to (Shah 2014). Spotify can also predict what songs are likely to make top of the charts; Spotify could not have achieved this success if they had worked alone.

The option of buying or partnering a bespoke IT company is not usually open to SMEs and they have to grow their own expertise. Making sense of data has always required appropriate skills and knowledge. There are particular problems associated with trying to explore large datasets, and visualization is critical – both for examining structures and identifying data problems. Errors in a large dataset may not be trivial to correct; indeed, the larger the dataset, the harder the task. One solution, in the UK at least, is to consider government-supported schemes, such as Knowledge Transfer Partnerships (KTPs). A government agency, called Innovate UK, part funds a graduate for one to three years to work on a well-defined project. The graduate is employed by a university and supervised by an academic but works full time at the company. KTPs are particularly suited to developing a new way of working as the KTP associate provides an independent, external viewpoint with a limited timescale. KTPs provide a low-risk way for a company to try a new venture such as integrated data analytics and develop new expertise in technical skills (Knowledge Transfer Partnerships 2015).

Another way to grow data-analysis skills is to nurture enthusiasm in young people, inspiring them to work in this area. An example is the Young Rewired State initiative in UK schools (http://yrs.io). This is now well established but was slow to get started. In 2009, Rewired State ran 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 (Rewired State 2015). 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 realized 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.”

Traditional Information Technology (IT) school teaching does not offer students the chance to really excel in innovation. Coding workshops, however, are a more appropriate way of engaging with young people. Each year Young Rewired State organizes a festival of code somewhere in the UK; entries to the final competition display a very high level of complexity and innovation.

The Digital Agenda of the EU, managed by the European Commission Directorate-General for Communications Networks, Content & Technology supports the Europe Code Week. This is a grassroots initiative which aims to bring coding and digital literacy to everybody in a fun and engaging way (http://codeweek.eu/). In October 2015, millions of children, parents, teachers, entrepreneurs, and policy makers came together in events and classrooms to learn programming and related skills. During the workshop, young people are asked to help test out state-of-the-art equipment such as 360° cameras and wrist-worn sensors, as well as learning about computational thinking by using LEGO.

3.2. *Lack of Statistical Skills*

The examples in Subsections 2.2.3 and 2.2.4 demonstrate the benefits of statistical analysis and data-mining techniques applied to integrated data. Basic descriptive analytics, as in Example 2.2.2, can also be very useful; however, these still require a certain level of confidence in handling numbers. A major reason for the lack of use of official statistics by business is poor numeracy and statistical competence within the business community. SMEs show little awareness of business statistics, as found in the BLUE-ETS study (Bavdaž 2011). To develop statistical literacy in the UK, the RSS launched *getstats* in 2010 to improve how the practical numbers of daily life, business and policy are handled (RSS 2015). Initiatives that support the ten-year campaign focus on three areas: the media, politicians and policymakers, and education, including higher education. The RSS is active in all these areas and has reached a broad spectrum of the population. Improvements in national statistical competence are difficult to quantify, but it is clear that acceptance of the importance of statistics has improved. For example, on World Statistics Day (October 20, 2015) the RSS delivered a training session to elected Members of Parliament in the London Parliament Building focusing on the use and interpretation of statistics in public life.

The increasing availability of Massive Open Online Courses (MOOCs) may help ease the difficulties encountered by SME staff wanting to acquire new skills and develop new learning. A wide range of courses is offered worldwide and staff have more choice about when they learn; see, for example www.mooc-list.com. Nevertheless, learning is not the same as doing, and it is important for staff to be allowed to experiment with company and official statistics data to see for themselves what benefits can be achieved.

Eurostat is the Statistical Office of the European Communities and as a Directorate-General of the European Commission is responsible for gathering data from NSIs throughout the European Community. Eurostat's mission is to provide the European Union with a high-quality statistical information service and to generate statistics at European level that enable comparisons between countries and regions. Eurostat does not collect data but consolidates it and has a large, harmonized cache of data accessible via its website (http.//ec.europa.eu/eurostat/). Eurostat recently showed its willingness to engage with users by providing a (free) training event to local government officers with sessions aimed at enabling businesses to access and utilize Eurostat data. The full attendance at the event indicates growing confidence in handling data and interest in data mining both at the descriptive and the predictive levels.

The wealth of official statistics is a source of fascination to academic statisticians, but few of them refer to such data in their lectures or use them in examples (Bavdaž 2011).

Clearly official statistics can provide excellent examples for data manipulation, tabulation, graphics, and statistical analysis. So, in addition to poor levels of numeracy, SME managers have little awareness of the free resources that would considerably help their businesses, and when they recall the statistics they have learned they do not immediately think of official statistics.

3.3. Lack of Interest

The examples in Section 2 derive from companies whose interest in data has been stimulated by personal contact with statisticians, by their susceptibility to articles in their trade press, and by their entrepreneurial spirit. This motivation has enabled them to overcome the typical challenges for SMEs when faced with the prospect of making use of business statistics. Many SMEs are discouraged by the lack of technically skilled staff, and by staff who are expected to multitask with little time available for proactively developing new expertise. They also worry about confidentiality issues and are wary of jeopardizing relationships with their customers. The main issue, however, is a general lack of interest. Nevertheless, as we found in an earlier research project focusing on helping SMEs adopt Six Sigma practices, there can be some sea changing improvements in performance in all parts of the business if SMEs are motivated by expert support, example, and peer-group activities bringing people together and giving them the chance to help each other (Stewardson and Coleman 2003). Becoming more data driven makes good business sense. As found in a survey reported in the Harvard Business Review (McAfee and Brynjolfsson 2012):

> "The more companies characterized themselves as data-driven, the better they performed on objective measures of financial and operational results. In particular, companies in the top third of their industry in the use of data-driven decision making were, on average, 5% more productive and 6% more profitable than their competitors."

It is essential to advertise the success of data analytics to inspire others, and suitable market places for such exchanges are beginning to emerge. The examples in Section 2 show some opportunities for SMEs gained by using official statistics. It is hoped that more case studies will become available and will be instrumental in encouraging other applications in this area. Joint projects often emerge when disparate groups come together. One of the reasons for the lack of use of official statistics is that data providers, data users, and entrepreneurs tend to be isolated from each other and do not often meet. Yet there have been some excellent examples of mutual benefit when these different groups cooperate and it is evidently important to bring people together to discuss their ideas.

Hacks are events in which data suppliers, data owners, designers, and IT specialists come together to analyze data using their combined expertise. The rationale is that mixing diverse skills will prompt new ways of thinking and create innovative solutions. For example, a Culture Code Hack held in Newcastle, UK, included librarians presenting data from lending libraries, arts managers with data on location and demographics of theatre goers, and musicians with songs from different parts of Northumberland. The author presented a large set of retail sales data from the ENBIS challenge (ENBIS 2015). Computer scientist hackers then spent the next 24 hours finding innovative ways to illustrate and interpret the datasets that particularly appealed to them. Groups then presented their work, which included interactive maps showing where and when people borrow computer software library books, what type of people go to which theatre performances and how far they travel, and what melodies to expect in different Northumberland villages. The outputs were fascinating and informative, but the lack of statistical thinking was disappointing. Statistical multivariate analysis and predictive modeling could have added substantially to the insight derived from the nevertheless excellent visualizations and user-friendly interfaces.

4. Discussion, Conclusions, and Recommendations

Businesses are becoming more interested in using their data more effectively, and amongst NSIs there is a determination to make their official statistics more widely appreciated in the business world. It is timely, therefore, to focus on how integrated data analytics can become more mainstream.

Networks and forums are an important way to help members share new ideas, and three examples are briefly described. The Association of Public Data Users (APDU 2015) in the US is a network that connects users, producers, and disseminators of government statistical data. The network members share interest in the collection, distribution, safeguarding, and explanation of public data. Organizations align themselves with APDU to gain a better understanding of public data, and potentially gain access to the public data for further analysis.

A recent cooperation between the ONS and the RSS resulted in the RSS web-based portal called StatsUserNet, which is aimed at general user engagement (StatsUserNet 2015). The interactive website provides a forum for all communities interested in official statistics. For example, the business and trade statistics community promotes dialogue, shares information and maintains close liaison between the producers and users of official business and trade statistics. Membership of individual communities of StatsUserNet has risen to over 3,000 people.

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 program and the BLUE-ETS research project that were initiated at approximately the same time, ENBES is endeavoring to bring business statistics closer to its users, as well as helping to gain a better understanding of user needs for business statistics. A seminar entitled "Statistics: About Businesses, for Businesses" was co-organized by ENBES, the RSS Official Statistics Section (OSS), and the RSS Quality Improvement Section (QIS) and held at the RSS in London. The OSS is interested in issues around collecting high-quality data and presenting them clearly and accurately, the QIS is interested in supporting statisticians using data to improve the effectiveness and profitability of enterprises in all sectors. This was the first time that the two sections had held a joint meeting and this reflects their growing interest in stimulating more effective use of official statistics, especially by SMEs. The seminar brought together practitioners, users, and methodologists, and addressed entrepreneurs, researchers, and policy makers. The example described in Subsection 2.2.4 above was presented as one of the case studies

(ENBES 2013). Such seminars are very useful in raising the profile of official statistics, showing their range and generating ideas on how they could be better communicated and utilized. A follow-up meeting was held at the RSS conference in Sheffield (ENBES 2014).

The initiatives for improving access to official statistics, such as Eurostat training, and making more data openly available are welcome mechanisms that encourage greater use of data. If it is made a requirement to publish reports from such open data analysis, there will soon be a copious supply of case studies from which businesses can learn. Statisticians should watch out for interesting case studies that either show evidence of statistical thinking or could be improved by statistical thinking and make sure that SMEs hear about them.

In conclusion, many companies are aware that they have valuable internal data but they do not maximize their use of them. They are receptive to being part of research initiatives aimed at helping them, providing they can see the direct benefits. Case studies are an important tool for increasing awareness. As the aim is to reach a wide audience, authors may be well advised to focus on publication in trade journals and the popular press, such as the RSS journal *Significance* and website “Stats Life”, as well as the website “Statistics Views” which aims to be a “one-stop shop” for the statistics community. This suggests the following recommendation:

- Statisticians need to publish case studies.

Future development of integrated data analytics requires that statisticians take part in entrepreneurial activities such as hacks. Coding sessions and events aimed at young people encourage the involvement of the next generation. These experiences help to overcome the barriers preventing fluent use of data and official statistics. Parents and teachers may not feel that they have time to spare to get involved in these activities, so it is very important for universities and businesses to offer as much support as possible, and statisticians are recommended to take an active part:

- Statisticians need to support hacks and coding events for young people.

The full attendance at Eurostat’s local government training indicates interest and confidence in handling data amongst regional policy makers and planners. What is now needed is a mechanism for cascading that enthusiasm from local government to the businesses they deal with. Statisticians should ensure that their business contacts, colleagues, and students are aware of networks such as ENBES and forums such as *StatsUserNet* which link government and business, and they are recommended to contribute case studies to encourage more extensive utilization of data:

- Statisticians need to help to cascade interest from government to businesses.

Researchers engaging in projects funded by the EU can help to promote the growth of integrated data analytics by SMEs. Projects can build on current initiatives such as the European Data Forum. Statisticians should be encouraged to include an exploration of official statistics in work packages in EU projects and KTP projects, providing a definite focus on the opportunities that they can bring. Academics should be encouraged to include projects that exploit the benefits of official statistics in Master’s-degree-level dissertations and doctoral training wherever possible. For example, Coleman (2014) described group project work which encourages integrated data analytics within the center for doctoral training in "Cloud computing for big data" at Newcastle University:

- Statisticians need to include integrated data analytics in EU projects, KTPs, and research training programs.

There are some excellent examples of open data integration; for example, Hans Rosling combines data on wealth and life expectancy in his video presentation "200 years that changed the world" (Gapminder 2015) and Stotesbury and Dorling (2015) integrate datasets on inequality and social outcomes, including environment, education, and health, and find some intriguing associations. However, neither of these examples focusses on the integration of open data and internal company data which is of particular promise for SMEs. Although some SMEs are making productive use of their data, carrying out data mining and making use of official statistics, it is unclear how prevalent this is. It is therefore recommended that:

- A survey of SME attitudes, activities, and needs should be carried out.

The interest in integrated data analytics shown by the companies in the examples in Section 2 was stimulated by contact with statisticians, by articles in their trade press and by their entrepreneurial spirit. Entrepreneurs are seeing the opportunities for adding value to open data and making business use of company data. Statisticians need to make a bigger impact both by active involvement in data communities and by writing wide-reaching articles. They need to keep pace with the changes taking place in their profession; they need to maintain contact with experts in other fields such as computer science, operations research, and data science to ensure that statistics continues to make a valuable contribution in the increasingly data-driven business world. Working alongside data owners, designers, hackers, and programmers, statisticians need to make sure that statistics as a subject is firmly established as an intrinsic component in helping businesses exploit the increasing availability of data.

5. References

Ahlemeyer-Stubbe, A. and S. Coleman. 2014. *A Practical Guide to Data Mining in Business and Industry*. London: Wiley.

APDU. 2015. The Association of Public Data Users. Available at: http://apdu.org/ (accessed August 2015).

Bavdaž, M. 2011. "Business Use of NSI Statistics Based on External Sources (NSIs, Publications, Expert Opinions)". Edited by M. Bavdaž. Available at: www.blue-ets.istat.it/fileadmin/deliverables/Deliverable3.1.pdf (accessed October 2015).

BIG. 2015. "BIG – Big Data Public Private Forum." Available at: http://www.big-project.eu/ (accessed August 2015).

BYTE. 2015a. "The Big Data Roadmap and Cross-DisciplinarY Community for Addressing SocieTal Externalities." Available at: http://byte-project.eu/ (accessed August 2015).

BYTE. 2015b. “Case Studies from Project.” Available at: http://byte-project.eu/wp-content/uploads/2015/06/FINAL_BYTE-D3-2-Case-studies-report-.pdf (accessed October 2015).

Coleman, S.Y. 2014. “Business and Academic Synergy in the World of Big Data.” In Proceedings of ENBIS-14 Conference of European Network of Business and Industrial Statistics, September 21–24, Linz, Austria. Abstract available at: http://www.enbis.org/activities/events/current/253_ENBIS_14_in_Linz//abstracts (accessed October 2016).

Coleman, S., S. Whitfield, J. Berry, G. Johnson, and P. Gore. 2015a. “Monetising Company Big Data.” In Proceedings of European Network of Business and Industrial Statistics (ENBIS) Spring Meeting on Predictive Analytics with Big and Complex Data, 2–5 June, Barcelona, Spain. Abstract available at: http://www.enbis.org/activities/events/current/379_ENBIS_Spring_Meeting_2015//abstracts. Project details at http://info.ktponline.org.uk/action/details/partnership.aspx?id=9671 (accessed October 2016).

Coleman, S., I. Zaman, R. Norman, and K. Pazouki. 2015b. “Business Strategy Embraces Data Analytics to Meet New Challenges in the Shipping Industry.” In Proceedings of ENBIS-15 Conference of European Network of Business and Industrial Statistics, 6–10 September, Prague, Czech Republic. Abstract available at: http://www.enbis.org/activities/events/current/380_ENBIS_15_in_Prague//abstracts. Project details at: http://info.ktponline.org.uk/action/details/partnership.aspx?id=9738 (accessed October 2015).

Coleman, S. and W. Yabsley. 2015. “Monetising Data in the Energy and Utility Sectors.” In Proceedings of ENBIS-15 Conference of European Network of Business and Industrial Statistics, 6–10 September, Prague, Czech Republic. Abstract available at: http://www.enbis.org/activities/events/current/380_ENBIS_15_in_Prague//abstracts (accessed October 2015).

Dalla Valle, L. 2014. “Official Statistics Data Integration Using Copulas.” *Quality Technology and Quantitative Management* 11: 111–131. Doi: http://dx.doi.org/10.1080/16843703.2014.11673329.

Dalla Valle, L. and R.S. Kenett. 2015. “Official Statistics Data Integration for Enhanced Information Quality.” *Quality and Reliability Engineering International* (advance online publication). Doi: http://dx.doi.org/10.1002/qre.1859.

Duedil. 2015. “Spotify Limited.” Available at: https://www.duedil.com/company/06436047/spotify-limited (accessed August 2015).

ENBES. 2013. Seminar announced at http://www.statisticsviews.com/details/event/4325891/Statistics-For-Businesses-About-Businesses-seminar.html (accessed October 2016).

ENBES. 2014. ENBES seminar presentations. Available at: http://www1.unece.org/stat/platform/display/ENBES/Session+on+Data+Collection+Challenges+in+Establishment+Surveys (accessed October 2016).

ENBIS. 2015. “ENBIS Challenge by JMP.” Available at: http://www.enbis.org/activities/jmp_challenge/index?_ts=85163 (accessed October 2015).

European Data Forum. 2015. *European Data Forum Homepage*. Available at: http://www.data-forum.eu/ (accessed August 2015).

European Data Forum. 2014. *European Data Forum (EDF)*. Available at: http://www.data-forum.eu/sites/default/files/EDF2014-Report.pdf (accessed August 2015).

Fontdecaba, S., P. Grima, L. Marco, L. Rodero, J.A. Sanchez-Espigares, I. Sole, X. Tort-Martorell, D. Demessence, V.M. de Pablo, and J. Zubelsu. 2012. "A Methodology to Model Water Demand Based on the Identification of Homogeneous Client Segments. Application to the City of Barcelona." *Water Resources Management* 26: 499–516. Doi: http://dx.doi.org/10.1007/s11269-011-9928-5.

Gapminder. 2015. *200 Years that Changed the World*. Video. Available at: http://www.gapminder.org/videos/200-years-that-changed-the-world/ (accessed October 2015).

Kenett, R.S. and G. Shmueli. 2016. Official Statistics. Chapter 10 in *Information Quality – The Potential of Data and Analytics to Generate Knowledge*, by R.S. Kenett and G. Shmueli: Wiley.

Knowledge Transfer Partnerships. 2015. *Innovate UK Homepage*. Available at: http://ktp.innovateuk.org/ (accessed October 2015).

McAfee, A. and E. Brynjolfsson. 2012. "Big Data: The Management Revolution." *Harvard Business Review*, October 2012. Available at: https://hbr.org/2012/10/big-data-the-management-revolution (accessed October 2015).

Mirani, L. 2015. "Why an Obscure British Data-Mining Company Is Worth $3 billion." Available at: http://qz.com/323944/why-an-obscure-british-data-mining-company-is-worth-3-billion/ (accessed August 2015).

ODI. 2015. *Open Data Institute Homepage*. Available at: http://www.theodi.org (accessed October 2015).

Research Councils UK. 2015. "Draft Concordat on Open Research Data." Available at: http://www.rcuk.ac.uk/research/opendata/ (accessed October 2015).

Rewired State. 2015. *Rewired State Homepage*. Available at: http://rewiredstate.org (accessed October 2015).

RSS. 2014. Data Manifesto. Available at: http://www.rss.org.uk/Images/PDF/influencing-change/rss-data-manifesto-2014.pdf (accessed October 2016).

RSS. 2015. "Statistical Literacy." Available at: http://www.rss.org.uk/RSS/Influencing_Change/Statistical_literacy/RSS/Influencing_Change/Statistical_literacy.aspx?hkey=821bf2f4-8a09-413c-8d22-290e2209a92a (accessed October 2016).

Shah, F. 2014. "How Spotify and Shazam Predict Music's Next Big Talent." Available at: http://dataconomy.com/predicting-musics-next-big-talent-big-data-case-spotify-shazam/ (accessed August 2015).

Snee, R.D. and R.W. Hoerl. 2004. *Six Sigma Beyond the Factory Floor: Deployment Strategies for Financial Services, Health Care, and the Rest of the Real Economy*. New Jersey: Prentice Hall.

StatsUserNet. 2015. "Business and Trade Statistics." Available at: http://www.statsusernet.org.uk/communities/viewcommunities/groupdetails/?CommunityKey=36dd28ed-e10a-440e-b7fb-86650b746c43 (accessed October 2015).

Stewardson, D.J. and S.Y. Coleman. 2003. "Success and Failure in Helping SMEs: A Three-Year Observational Study." *Industry and Higher Education* 17: 125–130.

Stotesbury, N. and D. Dorling. 2015. *Understanding Income Inequality and its Implications: Why Better Statistics Are Needed*. Statistics Views, October. Available at:

http://www.statisticsviews.com/details/feature/8493411/Understanding-Income-Inequality-and-its-Implications-Why-Better-Statistics-are-N.html (accessed October 2016).

UK Government. 2014. Press release: Government launches maritime study to boost growth. https://www.gov.uk/government/news/government-launches-maritime-study-to-boost-growth (accessed October 2016).

World Bank. 2015. "SME Statistics." Available at: http://siteresources.worldbank.org/CGCSRLP/Resources/SME_statistics.pdf (accessed October 2015).

Received March 2014
Revised October 2015
Accepted March 2016