BIG DATA FOR PRODUCT INNOVATION IN MANUFACTURING: EVIDENCE FROM A LARGE-SCALE SURVEY

Jasna PRESTER, Mihaela JURIĆ

Abstract: The article analyses big data usage in the Croatian manufacturing sector. Big data usage is still low but present. We analysed the influence of six sources of big data and their influence on share of returns generated by new products using two step OLS regression analysis. The results are robust but they show that some sources have positive and some have negative effects on share of returns generated by new products. Based on the most recent research of scholarly papers we define big data and show a clear research gap by linking big data and innovation. That is, only six papers deal with big data and innovation. In five papers big data comes from social media data, and in the remaining one paper they use data from sensors but predominantly to reduce cost or support the product. Therefore, we contribute by closing this research gap of linking big data and innovation.

Keywords: Big data; Croatia; EMS Survey; Manufacturing; Product innovation

1 INTRODUCTION

Big data is a buzz word that appeared approximately in 2005 according to three most recent scientific papers on big data literature research [1-3]. The current Google trends show even greater numbers than reported in [2, p. 97]. For the current state please see Fig. 1.

![Figure 1: Goggled: Innovation and Big data cumulative per year](image)

Big data is defined by 5V, namely; volume, variety, velocity, veracity, value [1-2]. However, for the purpose of this paper we adopted the description given by [2]: "extremely large amount of structured, semi structured or unstructured data continuously generated from diversified sources, which inundates business operations in real time and impacts on decision-making through mining insightful information from rambling data. For research clarity, big data includes large structured datasets and unstructured data in the form of text (e.g. documents, natural language), web data (e.g. web structure, web usage, web content), social media data (e.g. virtual network), multimedia data (e.g. image, audio, video), and mobile data (e.g. sensor, geographical location, application)." According to [2] who analysed more than 300 scholarly peer reviewed papers, only 6 papers deal with big data and innovation. We analysed all six of them: four of them use social data for innovation [4-7], one focuses the research on product defects discovery [8], and one uses social data for smart cities thus not directly related to product innovation [9]. Most papers, according to [2], focus on a specific problem that is solved by big data. Therefore, there is a clear gap in literature that is simultaneously using big data (but not social data) and innovation. We start the paper by explaining in more details what big data is, mostly based on three prominent literature reviews [1-3], which all investigated more than 150 scholarly journals. Then we briefly define innovation to a larger extent as defined by the OSLO manual [10]. In the methodology section, we describe the model, variables and data gathering methodology. Next, we present results and discussion. Finally, we conclude the article.

2 BIG DATA

The literature is abundant with grand terms of how big data will revolutionize innovation [11], the fourth industrial/scientific revolution [12], the next frontier for innovation [13], "transforming processes, altering corporate ecosystems, and facilitating innovation" [3]. However, as [3] states, potential adopters of ‘big data’ are struggling to better understand the concept and therefore capture the business value from ‘big data’. A recent report by McKinsey on big data shows the current state of big data in US, showing that manufacturing by far exceeds the data gathered when compared to other sectors [14]. This is illustrated in Fig. 2.

![Figure 2: McKinsey's Big Data and Innovation Comparison](image)

However, recent report by McKinsey [15] shows that despite the largest chunk of data being in manufacturing, only 20-30% of that data is actually used for improvement (p. 2). According to [2], a clear path towards management and usage of this data is an urgent need. But big data is not without challenges. Data has to be collected in a systematized way in order to be processed. The processing of data is a challenge
in itself [1]. So far, [1] identified the following processing methods: descriptive analytics scrutinizes data and information to define the current state of a business situation in a way that developments, patterns and exceptions become evident, in the form of producing standard reports, ad hoc reports, and alerts [16]; inquisitive analytics is about probing data to certify/reject business propositions, for example, analytical drill downs into data, statistical analysis, factor analysis [17]; predictive analytics is concerned with forecasting and statistical modelling to determine the future possibilities [18]; prescriptive analytics is about optimization and randomized testing to assess how businesses enhance their service levels while decreasing the expenses [16]; and preemptive analytics is about having the capacity to take precautionary actions on events that may undesirably influence the organizational performance, for example, identifying the possible perils and recommending mitigating strategies far ahead in time [19]. We overtook these definitions from [1] because all five data analysis processes would be beneficial for either innovation of the product or modifying a product based on the failure data recorded by either sensors or other data capturing techniques. This now brings to the question of what is defined as innovation.

According to Oslo manual (OECD 2005, 32) [10] product innovations are divided into improved products and products new to the market or radically new products. An improved product "is an existing product whose performance has been significantly enhanced or upgraded. A simple product may be improved (in terms of better performance or lower cost) through use of higher-performance components or materials, or a complex product which consists of a number of integrated technical sub-systems may be improved by partial changes to one of the sub-systems".

A new product "is a product whose technological characteristics or intended uses differ significantly from those of previously produced products. Such innovations can involve radically new technologies, can be based on combining existing technologies in new uses, or can be derived from the use of new knowledge".

In this analysis we define New products as both of these categories, even though in the questionnaire there is an additional question regarding these "Radically new products".

3 HYPOTHESES

According to the previous paragraph on innovation, it is evident that sources of idea coming from big data are not explored enough. Usual sources of ideas for innovation [10] (p. 78-80) can be internal sources of information (R&D, marketing, and production departments etc.) or external sources (customer/user, supplier, research units, conferences, scientific papers etc.). In this paper we investigate the sources of ideas coming from the usage of big data, which is stored in the manufacturing plant. Specifically, we explored data coming from providing remote support to customers, data coming from Sensor or remote control, data coming from the Enterprise Resource Planning software (ERP), data coming from exchange with supply chain partners (SCM), data coming from automation of flow of goods and storage, identification systems such as bar codes, RFID tags, etc., which for simplicity we abbreviated to (RFID), and data coming from digital devices used to program equipment which we abbreviated to Mobile programing. All these data sources fall into mobile big data as defined by [2]. All this data is supposed to enable companies to detect potential problems of the current product and enable and give ideas how to improve a product either incrementally giving raise to what Oslo manual calls Improved products or radically new products which Oslo manual defines as New products.

The model we propose is fairly simple. Each of these six sources will improve revenues from new products.

H1: data coming from Remote support will enhance share of revenues generated by new products

H2: data coming from Sensor or remote control will enhance share of revenues generated by new products

H3: data coming from ERP system will enhance share of revenues generated by new products

H4: data coming from SCM system will enhance share of revenues generated by new products

H5: data coming from RFID system will enhance share of revenues generated by new products
H6: data coming from Mobile programming system will enhance share of revenues generated by new products

These hypotheses will be tested through two step OLS regression analysis. Therefore, here is the place to introduce control variables.

4.1 Size of the Company

There is a difference in innovation output in small and large firms [26-28]. Bigger companies have larger and better R&D background, more staff, suppliers, customers that are all sources of innovative ideas. Size of a company is considered as a contingency because size of a company in terms of number of employees does not change overnight and depends also on labour market and overall conditions of the economy. Therefore, size of the company is considered as a control variable and a contingency; it is expected that larger companies will have more benefit in terms of generated revenues from new products. Therefore, H7 is as follows:

H7: Larger companies obtain higher share of return from new products

4.2 Complexity of the Product

Complexity is usually measured in number of components, newness, or number of functions designed into the product [29-31]. However, [32] in their research found that complexity also brings in new growth opportunities (58% of responders), and possible competitive advantage (59.4% of responders). [33] researched product complexity in new product development (NPD). Although, as [32] show, complexity of the product should increase new product potential developments, [33] found no impact of complexity on new product performance measures. The impact of complexity on manufacturing performance has not been clearly articulated in the previous empirical studies despite the widely expected negative relationship between them [34]. For example, [35] and [36] show that the higher the complexity of products is, the more complicated the supply chain is, and with that the risk of operating performance failures raises. [37] show that the more complex the product is, it might lead to poor delivery performance. [38] observe that the lead time increases with the number of parts. By analogy, complexity would also impact new product performance. Therefore, [32], [33] show that complexity might increase the chance of better innovative results, but also that it might bring problems to supply chain and consequently prolong the period of generating positive results. Therefore, we will hypothesize that H8 is as follows:

H8: complexity of the new product positively affects share of return from sales of new products

5 METHODOLOGY

The research data was collected using the European Manufacturing Survey (EMS), coordinated by the Fraunhofer Institute for Systems and Innovation Research – ISI, the largest European survey of manufacturing activities [39]. The survey’s questions deal with manufacturing strategies, application of innovative organizational and technological concepts in production, cooperation issues, production off-shoring, servitisation, and questions of personnel deployment and qualification. In addition, data on performance indicators such as productivity, flexibility, quality and returns are collected. The survey is conducted among manufacturing companies (NACE Revision 2 codes from 10 to 31) having at least 20 employees. The EMS project researches the whole manufacturing sector through a condensed eight-page questionnaire. To collect valid data permitting international comparisons, the EMS consortium employs various procedures recommended by the Survey Research Centre designed to avoid problems arising from the use of different languages and specific national terminology. First, a basic questionnaire is developed in English, which is then translated to the language of a country and then back to English to check consistency. Second, in each participating country pre-tests are conducted. Third, identical data harmonization processes are applied [40]. The sample used in the present paper consists of 105 Croatian manufacturing companies with over 20 employees. The questionnaire was sent to 1275 Croatian Chief Operating Officers who were asked to help in responding to the survey. A response rate of 8% was achieved, which is satisfactory for such large-scale voluntary surveys. The data collection was conducted in 2015.

6 RESULTS

We firstly analyse the sample using descriptive statistics. That is, we show the sample in terms of researched industries, size categories of companies and complexity of the produced product.

Distribution of industries, size of companies and complexity are given in Tab. 1 and Figs. 3-6.

Representativeness according to size and industry was performed and it was valid for both industry and size. That enables us to generalize conclusions for the whole Croatian manufacturing sector.

In the sample, 30.5% of companies are small with less than 50 employees, medium-sized companies having 50 to 249 employees are represented by 44.8% of companies, and 24.8% companies are large companies.

NACE code is not usually a good descriptor of complexity of the product, so additional analysis was performed in order to describe the sample in terms of complexity of the product they provide.

In the sample, 33.3% of companies produce simple products of not many parts, 41.2% of companies produce products of medium complexity, and 25.5% of companies produce complex products.

Fig. 5 shows share of revenues generated by new products depending on the complexity of the product.

From Fig. 5 it can be already seen that the Hypothesis H8 is confirmed, that more complex products can obtain higher share of revenues from new products.
Table 1 Distribution of the sample by industry

<table>
<thead>
<tr>
<th>No</th>
<th>Sector (NACE rev. 2) - 2 digits</th>
<th>Frequency</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>Manufacture of food products</td>
<td>11</td>
<td>10.5</td>
</tr>
<tr>
<td>11</td>
<td>Manufacture of beverages</td>
<td>2</td>
<td>1.9</td>
</tr>
<tr>
<td>12</td>
<td>Manufacture of tobacco products</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Manufacture of textiles</td>
<td>10</td>
<td>9.5</td>
</tr>
<tr>
<td>14</td>
<td>Manufacture of wearing apparel</td>
<td>2</td>
<td>1.9</td>
</tr>
<tr>
<td>15</td>
<td>Manufacture of leather and related products</td>
<td>3</td>
<td>2.9</td>
</tr>
<tr>
<td>16</td>
<td>Manufacture of wood and of products of wood and cork, except furniture; manufacture of articles of straw and plaiting materials</td>
<td>5</td>
<td>4.8</td>
</tr>
<tr>
<td>17</td>
<td>Manufacture of paper and paper products</td>
<td>4</td>
<td>3.8</td>
</tr>
<tr>
<td>18</td>
<td>Printing and reproduction of recorded media</td>
<td>5</td>
<td>4.8</td>
</tr>
<tr>
<td>19</td>
<td>Manufacture of coke and refined petroleum products</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>20</td>
<td>Manufacture of chemicals and chemical products</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>21</td>
<td>Manufacture of basic pharmaceutical products and pharmaceutical preparations</td>
<td>3</td>
<td>2.9</td>
</tr>
<tr>
<td>22</td>
<td>Manufacture of rubber and plastic products</td>
<td>8</td>
<td>7.6</td>
</tr>
<tr>
<td>23</td>
<td>Manufacture of other non-metallic mineral products</td>
<td>26</td>
<td>24.8</td>
</tr>
<tr>
<td>24</td>
<td>Manufacture of basic metals</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>25</td>
<td>Manufacture of fabricated metal products, except machinery and equipment</td>
<td>4</td>
<td>3.8</td>
</tr>
<tr>
<td>26</td>
<td>Manufacture of machinery and equipment n.e.c.</td>
<td>13</td>
<td>12.4</td>
</tr>
<tr>
<td>27</td>
<td>Manufacture of motor vehicles, trailers and semi-trailers</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>28</td>
<td>Manufacture of other transport equipment</td>
<td>2</td>
<td>1.9</td>
</tr>
<tr>
<td>29</td>
<td>Manufacture of furniture</td>
<td>4</td>
<td>3.8</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>105</td>
<td>100</td>
</tr>
</tbody>
</table>

Fig. 6 presents the percentage of companies using the digital source of data. The 7th column displays the question "Do you use your digital data" and it can be seen that 52% of companies do use their digital data. However, the question has limitations because it was not possible to ask for which purpose they use data (marketing, operations or like).

Tab. 2 presents the results of the regression analysis. Looking at overall results of the regression analysis $R = 0.526$ with significance $p = 0.024$ we can conclude that the model is significant. Coefficient of determinacy $R^2$ shows...
that 15.6% of share of revenues are generated by these sources of innovation ideas.

<table>
<thead>
<tr>
<th>Table 2 Results from the regression analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>Standardized Beta</td>
</tr>
<tr>
<td>-------------------</td>
</tr>
<tr>
<td>(Constant)</td>
</tr>
<tr>
<td>Number of employees</td>
</tr>
<tr>
<td>Complexity</td>
</tr>
<tr>
<td>NACE Code</td>
</tr>
<tr>
<td>Remote support</td>
</tr>
<tr>
<td>Sensor or remote control</td>
</tr>
<tr>
<td>ERP</td>
</tr>
<tr>
<td>SCM</td>
</tr>
<tr>
<td>RFID</td>
</tr>
<tr>
<td>Mobile programing</td>
</tr>
<tr>
<td>Data usage from digital sources</td>
</tr>
<tr>
<td>R</td>
</tr>
<tr>
<td>$R^2$</td>
</tr>
<tr>
<td>F</td>
</tr>
<tr>
<td>Sig.</td>
</tr>
</tbody>
</table>

*a Dependent Variable: Share of revenues by new products (%)*

Two step regression analysis first evaluates the influence of control variables. In our case that was size of the company, complexity of the product produced and NACE code. The control variables show that only complexity of the product matters and therefore the hypothesis H8 is confirmed. Hypothesis H7 that larger companies will have a higher share of revenues by new products is not confirmed, that is shares of revenues from the new products do not depend on size of the company, is not confirmed which is in line with [41], who show that even small companies with limited resources are able to successfully innovate. Control variables account to only $R = 0.229$, $R^2 = 5.2\%$, not significant $p = 0.170$, $F = 1.710$.

Contrary to our hypotheses, only usage of data from Enterprise Resource Planning software has most influence on revenues from new products. Although significance is over the limit of $p = 0.05$, this hypothesis H3 has to be rejected. Data coming from remote support H1, data coming from sensors and remote control H2, data coming from digital exchange of information from suppliers and customers H4 and data coming from identification devices (shortened RFID) all have to be rejected because they show a negative sign, i.e., this data for some reason negatively affects share of revenues generated by new products. However, results cannot be generalised as the significances are over the limit of $p = 0.05$. One hypothetical explanation might be that this data is still dominantly used for solving current operational problems and it has not yet been analysed as prescribed in [1] and thus this negative sign. If this data was analysed as prescribed by [1], then maybe the effects of analysing this data would be positive on share of revenues generated by new products. Even though this was not put in the form of a hypothesis, Data usage from digital sources has a small but positive, although not significant effect. However, even this percentage is low (only 52% of companies analyse their digital data), which is in line with McKinsey’s report [15] that although manufacturing generates and stores most data as compared to other industries (see Fig. 2), they are still not using it to their full potential.

To try to interpret these results we have to go back to the definition of "Big data" as given by [2]. They name sources of big data as Text, Web data, Social media data, Multimedia data and Mobile data. Of the five named sources of big data we have researched only the last source – mobile data (sensors, geographical location, and application). If one looks at Fig. 6, sensor and remote control source of data, one sees that this source of data is still very low (only 6% of companies gather data through this channel of data generation). Also, although the percentages are larger for sources of data from Enterprise Resource Planning and exchange of data with supply chain partners (SCM), the percentages are still very low (28%). It would be expected that in current days all of the companies, large and small, would have had installed an Enterprise Resource Planning system which is obviously still not the case (only 38% of companies use it). One possible explanation for this is that the research was conducted in 2015 and maybe the percentage in the next round, which is scheduled for this year (the survey is based on a three-year period), would be higher.

The limitation of the study is that, at the time the study was conducted, the question regarding what they use collected data for was not included in the questionnaire. This should be included in the next round as it may represent a source of competitive priority.

7 CONCLUSION

In this work we have shown on grounds of literature research that there is a clear gap in researching big data for usage for innovation. Also, most big data used in literature are Web data, Text data, and Social media data and only a small part of this big data is coming from mobile data (sensors, GPS or application). Therefore, in this work we have closed this gap by providing an analysis of how big data in Croatian manufacturing enhances or reduces share of revenues by new products as a measure of a successful innovation.

Of our 8 hypotheses only one has been partially confirmed (H3) – Enterprise Resource Planning systems positively affect share of revenues generated by new products, and H8 has been fully confirmed (Producers of complex products obtain better share of revenues from new products), while other hypotheses had to be rejected. They have shown a negative sign, contrary to our hypotheses, although the coefficients were not significant at $p=0.05$. Based on descriptive data we hypothesized that the rejection of these hypotheses is for one in low level of usage of big data, and it is probably more used for problem solving than for analysing this data for new potential improvements and new products. This is actually in line with current research that manufacturing is still not using the full potential of the data it gathers as suggested in McKinsey’s report [15].

Some general conclusion and advice to managers is to invest more into some kind of ERP system and to analyse this data, as those two sources showed positive effect on share of revenues generated by new product. As for the other four sources of data (Remote support, Sensors and Remote control, Supply Chain Management Software and
Identification devices (abbreviated RFID), it might be useful to analyse the data after the problem is solved. As it seems now, the data is used for solving operational problems and maybe not enough effort is put into post analysis of this data as described in McKinsey’s report [15].

Acknowledgement

This research is conducted by Grant O-1861-204-3535 Building Competitiveness of Croatian Manufacturing financed by Croatian Science Foundation.

8 REFERENCES

https://doi.org/10.1111/1540-5885.1130236

https://doi.org/10.1287/mnsc.47.1.189.10662

https://doi.org/10.1142/S0219877012500435

https://doi.org/10.1080/09537287.2015.1027318

https://doi.org/10.1016/j.ijpe.2015.07.036


https://doi.org/10.1109/TEM.2002.803387


https://doi.org/10.1108/JOCM-04-2013-0052

https://doi.org/10.1111/jpim.12373

Authors’ contacts:

Jasna PRESTER, PhD, Professor
University of Zagreb,
Faculty of Business & Economics,
Trg J. F. Kennedy 6, 10000 Zagreb, Croatia
jprester@efzg.hr

Mihaela JURIĆ, univ. spec. oec., PhD candidate
University of Rijeka,
Faculty of Economics and Business,
Infodom d.o.o.
Ul. Andrije Žaje 61, 10000 Zagreb, Croatia
mihaela.juric@infodom.hr