Predictive Analytics for Supply Chains: a Systematic Literature Review

Kevin Bonnes
University of Twente
P.O. Box 217, 7500AE Enschede
The Netherlands
K.G.Bonnes@student.utwente.nl

ABSTRACT

Predictive analytics is a group of methods that uses statistical and other empirical techniques to predict future events, based on past occurrences. Predictive analytics can generate valuable information for the management of a supply chain company to improve decision-making. Even though the importance of the topic is clear, there is no clear overview of the use of predictive analytics in the supply chain currently. This research provides a state-of-the-art by performing a systematic literature review. In the literature review we have found the models, methods, techniques and applications of predictive analytics in the supply chain and have determined the trends and literature gaps. The most important finding is that even though there is only a limited amount of literature available, the interest in this topic is growing gradually. We also provide future research directions for further research on this subject.

Keywords
Literature Review, Predictive Analytics, Prediction, Supply Chain, Big Data, Forecasting

1. INTRODUCTION

Predictive analytics is a group of methods that uses statistical and other empirical techniques to predict future events, based on past occurrences. Although predictive analytics has become a hot topic recently, there are many companies within the supply chain sector that don’t make use of it yet [61]. As literature states, enhancing effectiveness and efficiency of supply chain analytics is a critical component of a company’s ability to achieve competitive advantage [61]. Predictive analytics for supply chains is therefore an interesting topic.

Predictive analytics are used in many different fields. This research focuses on the supply chain sector. An example of the use of predictive analytics in the supply chain is predicting the arrival times of a truck or a ship. There are multiple variables that can influence arrival times, for example: the weather, traffic congestion and the driver’s driving style. If the reliability of predicting arrival times can be increased, costs can be saved within the entire supply chain. These costs can be found in waiting times for trucks in distribution centers and improved decision-making.

If companies get a better understanding of how predictive analytics should be used, decision-making on management level will be improved. Currently there is literature available on this research area but there is no overview. This research will provide this overview by mapping out all of the relevant information available by conducting a systematic literature review, as described by Webster & Watson [78]. Afterwards, the literature will be categorized and presented in terms of models, methods, techniques and applications. We will also show the current trends in predictive analytics for the supply chain, to see where the field is going.

1.1 Problem statement

Even though there is a large body of work available on both predictive analytics and the supply chain, there has not been an overview of the combination of these subjects. This literature study will fill that gap by providing insight on the current use of predictive analytics for supply chains.

1.2 Research questions

Based on the problem statement, we arrive at the following research questions:

RQ1: What are the models, methods, techniques and applications used in predictive analytics for the supply chain?

RQ2: What are the research gaps of the literature found?

1.3 Remainder of this paper

The remainder of this paper is organized as follows: Section 2 describes the research method. In section 3 the results are presented. Section 4 provides a discussion of the results, the current trends and the gaps in the literature. Section 5 presents the conclusions, limitations and provides future research directions.

2. RESEARCH METHOD

In order to answer the research questions and validate the findings a systematic literature review was performed, following the method described by Webster & Watson [78]. This method was chosen to increase the scientific value of the paper by making sure that all of the relevant literature on this subject is taken into account. In this section each step of the literature search will be described and motivated so that our findings can easily be replicated and validated. In order to keep the quality of the research as high as possible our literature search was limited to only those databases, journals and conferences that have a good academic reputation.
Literature was searched in the databases Scopus and Web of Science. These are two of the leading databases and provide good coverage of the relevant sources.

The next step in the process was carefully selecting the journals and conferences, to make sure that we only include the relevant literature and don’t exclude any relevant paper by omitting a quality publication. Relevant journals were determined by creating a list of relevant research areas and selecting the top journals in each of these research areas. A list of the research areas can be found in Appendix A.

For each of the research areas a list of journals was determined based on the top 20 journals ranked by impact factor, from the ISI Journal Citation Ranking [72]. Irrelevant journals (based on topic) and duplicate journals were excluded. The list of journals was verified and validated by several existing journal rankings [4, 23, 46, 47, 51, 63, 64, 69, 71]. For each of these rankings the top 20 was compared with the ISI JCR top 20 and missing journals were added. This was done to make sure that we did not miss out on any important journal in any of the research areas. Eventually, a list of 137 journals was created. The journal list and the validation can be found in Appendix B.

Conference papers are generally of less importance compared to journals in the business/SCM and information systems disciplines, except for the computer science research area [73]. Conference papers however have a shorter throughput time and are thus often more recent and provide a better state-of-the-art than journal papers [28]. Hence, a total of 51 important conferences for the computer science and information systems area were also reviewed. A list of the selected conferences can be found in Appendix C. This list was based on several conference rankings by field rating [33, 36, 48]. To ensure the novelty of the conference papers found we specified an extra criterion: papers have to be published in 2010 or later.

Based on the research questions a research query was formulated. The research query was formulated in several steps of improvement, to make sure we did not miss or exclude any relevant papers. The research query can be found in Appendix D. To increase the likeliness of the results being relevant, the keywords in the research query have to occur in the titles, abstracts or keywords of the articles.

With the research query and the journal-conference-query, a literature search was performed. In Scopus, a total of 446 articles were found. In Web of Science, we found a total of 429 articles. After removing the duplicate articles that were found in both Scopus and WoS, a total of 581 papers remained.

3. RESULTS

The results of the literature search were evaluated on abstracts by two reviewers, to make sure that the papers that had no relation with both the supply chain and predictive analytics were excluded. After the selection process a total of 160 papers remained. The remaining set of papers was read and evaluated more in depth, until a final set of 64 papers was composed. This was done on the following criterion: Does the paper formulate a clear model that can be used for predictive analytics in the supply chain sector?

Based on the final collection of papers, a concept-centric literature review was performed. A concept-centric literature review is a literature review in which the concepts determine the structure of the review rather than the authors [78]. This means the results will be categorized in terms of concepts. Categorizing the literature plays an important role as it gives a good overview of the available literature and the literature gaps. Categorization is done on four different aspects of the literature found: models, methods, techniques and applications. These four aspects have been chosen because we think they are the most valuable as they give a good understanding of why and how predictive analytics should be used.

3.1 Models

Models in predictive analytics for supply chains can be divided into two categories: explanatory models and predictive models [66, 67]. Explanatory models (often statistical models) test a causal hypothesis based on historical data, and provide a conclusion about the model. They do not use new data to make predictions. Predictive models are also based on historical data, but make use of new data to test the model and make predictions about the future. Of models found in the 64 papers reviewed, only 18 were categorized as predictive models. Even though explanatory models do not imply predictive power [67], we do take them into account mainly because of a lack of ‘real’ predictive models. An overview can be found in Table 1.

<table>
<thead>
<tr>
<th>Type of model</th>
<th>Amount of papers</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Explanatory</td>
<td>46x (72%)</td>
<td>[3, 6–10, 12–14, 16, 18–22, 24–27, 30, 32, 37, 39–43, 45, 52–54, 56, 58–60, 62, 65, 70, 76, 77, 79–84]</td>
</tr>
<tr>
<td>Predictive</td>
<td>18x (28%)</td>
<td>[1, 2, 5, 11, 15, 17, 19, 20–22, 26, 27, 29–32, 34, 35, 37–45, 49, 50, 52, 57, 59, 60, 62, 65, 68, 70, 75–77, 79–84]</td>
</tr>
</tbody>
</table>

3.2 Methods

The methods used in the literature are made up of two parts: the data collection methods, either empirical or simulation, and the analysis techniques which will be discussed in the next section.

Data collection methods are the methods used to collect data as an input for the model. We categorized the data collection method in two different types: methods using empirical, measured data and methods using generated data. An overview of the data collection methods can be found in Table 2.

<table>
<thead>
<tr>
<th>Category</th>
<th>Methods</th>
<th>Amount of papers</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Generated data</td>
<td>Simulation</td>
<td>11x (17%)</td>
<td>[1, 2, 16, 19, 24, 25, 53–56, 58]</td>
</tr>
</tbody>
</table>
3.3 Techniques

Analysis techniques (e.g. time series, regression), also often referred to as research methods, are the underlying techniques used to analyze the data. Models in predictive analytics are always based on one or more techniques. An overview of the techniques used in the literature can be found in Table 3. In Table 4, the top 5 techniques are shown with reference to the models.

<table>
<thead>
<tr>
<th>Table 3. Techniques</th>
</tr>
</thead>
<tbody>
<tr>
<td>Technique</td>
</tr>
<tr>
<td>Time series</td>
</tr>
<tr>
<td>Regression</td>
</tr>
<tr>
<td>Neural networks</td>
</tr>
<tr>
<td>Mixed integer programming</td>
</tr>
<tr>
<td>Exponential smoothing</td>
</tr>
<tr>
<td>Fuzzy logic</td>
</tr>
<tr>
<td>Simulation</td>
</tr>
<tr>
<td>Statistical techniques</td>
</tr>
<tr>
<td>Mathematical techniques</td>
</tr>
<tr>
<td>Structural equation modelling</td>
</tr>
<tr>
<td>Support vector machines</td>
</tr>
<tr>
<td>System dynamics</td>
</tr>
<tr>
<td>Partial least squares</td>
</tr>
<tr>
<td>Logistic regression</td>
</tr>
<tr>
<td>Linear regression</td>
</tr>
<tr>
<td>Markov prediction</td>
</tr>
<tr>
<td>Markov decision processes</td>
</tr>
<tr>
<td>Markov chain method</td>
</tr>
<tr>
<td>Nonparametric regression</td>
</tr>
<tr>
<td>Continuum approximation</td>
</tr>
<tr>
<td>Analytic hierarchy process</td>
</tr>
<tr>
<td>Rough set theory approach</td>
</tr>
<tr>
<td>Machine learning</td>
</tr>
</tbody>
</table>

3.4 Applications

The applications of predictive analytics for supply chains are also categorized. These can be found in Table 5. This table gives an overview of the usage of predictive analytics within the supply chain.

As shown in Table 6, there are three major applications of predictive analytics within the supply chain: forecasting, inter-organizational collaboration and monitoring. In Table 6 for each of these applications the amount of papers found is shown with reference to the two types of models.

<table>
<thead>
<tr>
<th>Table 4. Techniques vs. Models</th>
</tr>
</thead>
<tbody>
<tr>
<td>Technique</td>
</tr>
<tr>
<td>Time series</td>
</tr>
<tr>
<td>Regression</td>
</tr>
<tr>
<td>Neural networks</td>
</tr>
<tr>
<td>Mixed integer programming</td>
</tr>
<tr>
<td>Exponential smoothing</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 6. Applications vs. Models</th>
</tr>
</thead>
<tbody>
<tr>
<td>Application</td>
</tr>
<tr>
<td>Forecasting</td>
</tr>
<tr>
<td>Inter-organizational collaboration</td>
</tr>
<tr>
<td>Monitoring</td>
</tr>
</tbody>
</table>

3.5 Trends

Besides the tables presented above, 4 line charts were generated to be able to detect trends. The line charts show the amount of relevant papers for each year, for both explanatory models and predictive models. Figure 1 presents a general chart and Figure 2, 3 and 4 present line charts respectively for each of the applications forecasting, inter-organizational collaboration and monitoring.
<table>
<thead>
<tr>
<th>Category</th>
<th>Topic(s)</th>
<th>Amount of papers</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Forecasting</td>
<td>Demand/order forecasting, inventory forecasting, predicting bullwhip effect, scheduling, predicting production volatility</td>
<td>34x (51%)</td>
<td>[1, 3, 9, 12–14, 17, 19–21, 29, 31, 35, 37, 41, 43, 45, 49, 52–54, 56, 58–60, 62, 70, 75–77, 79, 81–83]</td>
</tr>
<tr>
<td>Inter-organizational collaboration</td>
<td>Predicting influence of collaboration, supplier selection, predicting supply chain integration, predicting stakeholder collaboration strategies</td>
<td>13x (19%)</td>
<td>[5, 8, 13, 15, 18, 22, 27, 30, 32, 42, 57, 65, 84]</td>
</tr>
<tr>
<td>Monitoring</td>
<td>Container tracking, travel time prediction, logistics planning, predicting order arrival, predicting flow times, predicting RFID performance, inventory tracking, predicting logistic costs</td>
<td>11x (16%)</td>
<td>[6, 11, 16, 25, 26, 34, 39, 44, 55, 58, 68]</td>
</tr>
<tr>
<td>Price forecasting</td>
<td>Price forecasting</td>
<td>2x (3%)</td>
<td>[13, 38]</td>
</tr>
<tr>
<td>Product deficiency</td>
<td>Error prediction, predicting product failure</td>
<td>2x (3%)</td>
<td>[2, 10]</td>
</tr>
<tr>
<td>Other</td>
<td>Predicting supply chain performance, predicting cost-to-serve for new customers, predicting price negotiation outcomes, predicting customer purchases, predicting critical factors in a supply chain</td>
<td>5x (7%)</td>
<td>[7, 24, 40, 50, 80]</td>
</tr>
</tbody>
</table>

Figure 1. Models vs. Time

Figure 2. Models vs. Time, Forecasting
4. DISCUSSION

4.1 Models
As shown in Table 1, out of the 64 papers only 18 of them included a predictive model. This means only 18 papers actually used new data to make predictions about the future based on the model. The other papers presented explanatory models which can be used to make predictions but did not directly apply the model to new data. This means the 46 papers that presented an explanatory model did not directly make use of predictive analytics, but instead only presented a model, which implies that true analytics are scarce in the supply chain field.

4.2 Methods
Data collection methods in predictive analytics are mainly focused on empirical, measured data. As shown in Table 2, 53 of the 64 papers make use of measured data to test their model. Measured data consists of empirical or industry data, often generated by methods such as experiments, case studies or surveys, or a combination of these methods. Simulated data is data generated by simulations. Simulated data is often used due to a lack of measured data. As expected the percentage of papers using simulated data as input is quite low compared to the percentage of papers using measured data.

4.3 Techniques
As shown in Table 3, predictive analytics consists out of multiple different techniques. Some models use a combination of techniques (e.g. [6, 20, 60]). By far the most used technique is time series. Time series techniques are used to identify a pattern in a sequence of data points. Other widely used techniques are regression, neural networks, exponential smoothing (a technique used in combination with time series) and Mixed integer (non-)linear programming.

In Table 4, the techniques are shown with reference to the two types of models. As shown, the most used technique, time series, is almost always used in explanatory models. The percentage of papers using time series techniques in a predictive model is only 7%. This number is quite low compared to the average of 28% of predictive models in general (Table 1). Besides time series techniques, this also applies to regression techniques and exponential smoothing techniques. The only technique within the top 5 that is being used in an exceptionally high percentage of predictive models compared to explanatory models is mixed integer programming. In general, the percentage of predictive models compared to explanatory models for the top 5 techniques is only 19%, compared to the 28% of Table 1.

4.4 Applications
For applications in predictive analytics for the supply chain there are three major categories: forecasting, inter-organizational collaboration and monitoring. 86% of the papers can be classified in one of these three categories.

The first category, forecasting, encompasses several topics such as demand forecasting, inventory forecasting and predicting the bullwhip effect. Demand forecasting (e.g. [14, 43, 83]) is the forecasting activity in which a company tries to predict the quantity of products the customer will purchase. Predictive analytics are used to predict future demands based on historical demand information. Demand forecasting is often done by applying time series to a given set of input data. Inventory forecasting is closely related to demand forecasting. With inventory forecasting a company tries to predict the amount of inventory required to fulfill customer orders. Examples of papers about inventory forecasting are [59, 77].

Predicting the bullwhip effect is also related to demand forecasting. The bullwhip effect is the problem that arises due to a continuously changing demand of orders, which makes it hard to predict the amount of produced products. Examples of papers presenting models to predict the occurrence of the bullwhip effect are [53, 54, 56].

The second category is the inter-organizational collaboration applications. Examples of inter-organizational collaboration applications are supplier selection and predicting the influence of collaboration on supply chain performance. Supplier selection is about predicting the best possible supplier based on historical information about similar situations. Examples of papers providing a model for optimal supplier selection are [27, 30, 42].

Predicting the influence of supply chain collaboration is about sharing information between supply chain partners to improve the overall performance of the supply chain. Examples of papers are [5, 8].
The third category is monitoring applications. Examples of monitoring applications are travel time prediction and order arrival time prediction. Both of these applications are closely related. Travel time prediction is often used to predict order arrival times. Predictive analytics in this application are used to predict travel times based on historical information of similar routes, taking into account the influence of external factors such as the weather and traffic congestion. Examples of papers about travel time prediction or arrival time prediction are [11, 68]. Another example of monitoring is predicting transport costs. Examples of papers are [16, 55].

As shown in Table 5, a lot of research is done in the forecasting application compared to the other applications. However, most of the research done in the forecasting application only provides an explanatory model, as Table 6 shows. There is a limited amount of research that actually provides predictions. On the other hand, in the monitoring application field the amount of predictive models is relatively high.

4.5 Trends
Another aspect of our literature study is determining the trends of predictive analytics for supply chains. Figures 1 to 4 show the amount of papers published per year per application, together with their trend lines. There are a number of conclusions that can be drawn from these four figures.

First of all, as shown in Figures 1 to 4, predictive analytics for the supply chain is still a relatively premature research area. There is a limited amount of research available on this subject and most of the research was performed in the past few years. As shown in Figure 1, trend lines for both explanatory modeling and predictive modeling are increasing, slowly but gradually. Trend lines would increase even more if the year 2014 was removed from the figures, as 2014 shows a big drop in the number of papers published due to many papers not having been published yet.

Second, trend lines in explanatory models seem to increase faster than the trend lines in predictive models. Trend lines for predictive models are often more flat, except for the trend line in Figure 3, which can be ignored due to a low amount of papers in this application. This implies that over time, the percentage of explanatory models compared to predictive models increases. However, as shown in Figure 1, the increase of explanatory models is only slightly. Therefore it’s difficult to predict in which direction the research will be going in the future.

Another point of interest is that predictive models are slightly more recent than explanatory models. The first paper about predictive models for predictive analytics in the supply chain sector was released in 2004. In 2011 there was a relatively big increase in papers regarding predictive models. This could be explained by the fact that Shmueli [66] released a paper about the difference between predictive models and explanatory models in 2010. This can be an indication that predictive models are still emerging.

Finally, research on both the applications monitoring and inter-organizational collaboration started late, compared to forecasting. This applies to both papers about predictive models and papers about explanatory models. However, even though forecasting seems to be an older research area there is still an upward trend going on in this area. It is to be expected that the amount of research to all three of the applications will continue to increase in the future.

4.6 Literature gaps
As mentioned earlier, literature regarding predictive analytics for supply chains is still in its infancy. There are several literature gaps identified from the current literature. These literature gaps will be discussed on a high level, for each of the categories discussed earlier: models, methods, techniques and applications.

4.6.1 Models
There is a low amount of papers available on predictive analytics actually giving predictions. Most of the papers present a model, but do not apply new data to this model. This is a major gap that does not only apply to predictive analytics for the supply chain, but also for predictive analytics in general [60].

4.6.2 Methods
Regarding the data collection methods, there are a low amount of models based on data generated by simulation as expected. It is unknown however if there are any consequences regarding using simulated data or measured data.

4.6.3 Techniques
As shown in Table 3, well known techniques such as machine learning, support vector machines and system dynamics are clearly under-represented. This also applies to other techniques such as naive Bayes, k-nearest neighbours and decision trees, which are not even present at all. It would be interesting to see if these techniques can also be used in predictive analytics for the supply chain. This could also cause new trends to appear in this research area.

4.6.4 Applications
Besides the three major applications of predictive analytics for supply chains, predictive analytics can also be used for other purposes. Under-researched applications as become clear from Table 5 are price forecasting and predicting product deficiency. Successfully applying predictive analytics to one of these applications can be a cost-saving process.

Besides the two mentioned applications above, there is also a lack of research on customer sentiment data and the usage of predictive analytics in human resources [74]. Data can be gathered in both of these areas and thus can be potentially interesting for applying predictive analytics.

5. CONCLUSION

RQ1: What are the models, methods, techniques and applications used in predictive analytics for the supply chain?

There is only a limited amount of research available on predictive analytics for supply chains. The relevant literature was mostly released in the past few years but is growing rapidly. It is expected that the amount of literature released on a yearly basis will continue increasing the upcoming years.

The literature was categorized based on models, methods, techniques and applications. As shown in Table 1, the majority of the models are explanatory. Explanatory models provide a model based on historical data but do not apply new data to make predictions about the future, in contrary to predictive models.

Data collection methods found were mostly based on measured data. These data collection methods include experiments, case studies and surveys.

The most used techniques in predictive analytics for supply chains are time series, regression and neural networks.
However, the usage of the most used techniques in predictive models seems to be lower than the usage of less used techniques in predictive models, percentage-wise.

Applications of predictive analytics for supply chains are mostly found in the categories forecasting (e.g. demand forecasting, inventory forecasting, predicting bullwhip effect), inter-organizational collaboration (e.g. supplier selection, predicting the influence of collaboration) and monitoring (e.g. travel time prediction, order arrival time prediction, predicting logistic costs).

Based on the categorization, some trends were identified. As shown in the figures, the amount of papers on predictive analytics for the supply chain are still increasing. The number of papers about explanatory models seem to increase faster than the percentage of papers about predictive models however, but papers about predictive models are slightly more recent on average.

Research on the applications, monitoring and inter-organizational collaboration seem to be newer compared to forecasting. The forecasting application is more mature compared to the others, both based on time and on quantity.

**RQ2: What are the research gaps of the literature found?** We have also identified research gaps. First, there is only a low amount of papers that are actually making predictions. Most of the papers discuss an explanatory model instead of a predictive model. Second, there is a lack of research in the applications other than the three mentioned before. These applications include: price forecasting and predicting product deficiency.

5.1 Limitations and future research

A major limitation is the lack of papers on this subject. As mentioned earlier and shown in Figures 3 and 4, it is hard to predict trends based on only a few papers on a specific application. The reliability of some of the trends is therefore questionable. It is expectable however that research on this subject will increase as the trends show. Future research could therefore provide a better overview of the current state-of-the-art with an increase of relevant papers.

6. ACKNOWLEDGEMENTS

A special thanks to supervisor Sjoerd van der Spoel for providing feedback and assistance during the research.

References


**APPENDIX**

**A. LIST OF RESEARCH AREAS**

- Information Systems
- Computer Science: Artificial Intelligence
- Computer Science: Interdisciplinary Applications
- Engineering: Industrial
- Operations Research
- Supply Chain
- Decision Sciences
- Business
- Management

**B. LIST OF SELECTED JOURNALS**

The list of journals included can be found with the following link: https://docs.google.com/spreadsheet/ccc?key=0AhDiBPUDmbF1GB0O5IWUdhRXpfdmxMzlGaNVT3c&usp=sharing

The table shows the journal titles as well as their appearance in the journal rankings.

**C. LIST OF SELECTED CONFERENCES**

The list of the conferences included can be found with the following link: https://docs.google.com/spreadsheet/ccc?key=0AhDiBPUDmbF1GB0O5IWUdhRXpfdmxMzlGaNVT3c&usp=sharing

The table shows the conference titles as well as their appearance in the conference rankings.

**D. RESEARCH QUERY**

predict* AND
(“supply chain” OR SCM OR logistics) AND
(analy* OR “big data” OR “data mining” OR “business intelligence” OR “decision support”)