The role of domain analysis in prediction instrument development

Sjoerd van der Spoel – Chintan Amrit – Jos van Hillegersberg

University of Twente
7500 AE Enschede
The Netherlands
{s.j.vanderspoel,c.amrit,j.vanhillegersberg}@utwente.nl

ABSTRACT: *In order to develop prediction instruments that have sufficient predictive power, it is essential to understand the specific domain the prediction instrument is developed for. This domain analysis is especially important for domains where human behavior, politics, or other soft factors play a role. If these are not well understood, the predictive power of the prediction instrument would be severely affected.*

In this paper, we provide literature based reasons for the use of domain analysis for the development of prediction instruments, and we discuss the circumstances under which domain analysis is especially important. We present a structured literature review of the actual adoption of domain analysis for predictive analytics. That shows that few papers discuss how domain analysis was performed, and when it is discussed, the type of analysis often does not fit with the type of domain. As these papers do show adequate predictive power, we believe that the domain analysis in these papers was done implicitly.

To make the process of prediction instrument development, including domain analysis more transparent, we present requirements for a method for prediction instrument development, and an outline for such a method based on those requirements.

KEY WORDS: *Domain analysis, prediction instrument development, domain complexity*

1. Introduction

Variable selection and data cleaning are key steps in the creation of prediction instruments (Shmueli, 2010; Shmueli & Koppius, 2011). Together, they from a major part of understanding the domain the prediction instrument is developed for. This often depends on the expertise of the prediction instrument developer. Current research on prediction instrument development is often opaque about its approach to domain analysis, and it is unclear what can be done in situations where the analyst does not have sufficient expertise. In this paper, we bridge this gap by investigating the role of domain analysis in prediction instrument development. In particular, we investigate how domain analysis should be executed depending on the type of domain. We use the term *prediction instrument* throughout this paper, rather than the often used term predictive *model*, as a predictive *model* only implies a single model produced by a data mining or statistical technique, such as a decision tree model. Such a model does not contain how data should be cleaned before it is fed into the model, or which variables should be used to create new predictive models. Therefore, we use the term prediction instrument for a combination of a data selection & cleaning strategy and a technique or algorithm, that itself can make new predictive models.

The paper is structured as follows. Section 2 addresses reasons for domain analysis & circumstances under which domain analysis is especially important. Section 3 presents a structured literature review to understand how domain analysis is treated in existing research. Section 4 presents both requirements for a method for domain analysis, and an outline of a method that fits those requirements.

2. Theoretical Background

Prediction instruments are developed within a particular context, such as an organization or group of organizations. This context, which is a combination of several parts that have emergent properties, is referred to as a *system* (Checkland, 1989) or as a *domain*. The term ‘domain’, as used in this research also includes the system - it is everything that *influences* or *interacts* with the prediction instrument under development.

To make predictions, a prediction instrument requires data. This is either used to train a predictive model, or it is a new observation for which the outcome is to be predicted. This data is provided by/recorded in the domain. This means that (relevant aspects) of the domain are *represented* by the data. Key to domain analysis for prediction instrument development are understanding the domain and understanding how it is filtered through into data. This determines, amongst others, what variables to select and how to perform data cleaning, which are both important steps in predictive model development methodologies such as (Shmueli & Koppius, 2011) or CRISP-DM (Chapman et al., 2000). However, neither method clarifies how this domain understanding should be undertaken.

Existing literature describes roughly two approaches to domain & filter understanding. The first, *systems thinking*, considers each domain as an instance of an abstract, well-known domain (Checkland, 1989; Forrester, 1994). The understanding of that abstract domain can then be immediately transferred to the instantiation.

The second group of approaches, which contains *soft systems methodology* (Checkland, 1989) and *system dynamics* (Forrester, 1994) argues that this way of thinking is not applicable to domains where aspects such as human behavior play a role (Checkland, 1989; Forrester, 1994). The argument being that these are too different from an abstracted variant.

The systems where the aspects above play a role are called *complex domains*. Jackson & Keys (1984) provide the following definition: A complex domain is a domain where people are trying to act together; that involves politics, myths and meanings; that is unstructured; that is not fully observable; where parts have their own goals; that is probabilistic in nature; open to its environment; and subject to behavioral influences (Jackson & Keys, 1984).

In a complex domain, the filter between domain and data could be affected, adding noise or bias to the data. Data might be manipulated to be more favorable to certain actors in the domain, or mistakes could have been made in recording data, as previous work has shown (van der Spoel, Amrit, & van Hillegersberg, 2015, 2013).

The 'systems thinking' approach to domain analysis is what we refer to as *soft-exclusive domain analysis*, as in abstracting the domain, soft factors such as behavior, politics and individual goals are left out. Likewise, approaches that do consider these factors are *soft-inclusive domain analysis*. Based on Checkland & Forrester's (1989) work, soft-inclusive analysis should be used for complex domains. For simple domains, soft-exclusive analysis is the best fit.

For a soft-exclusive domain analysis, a good candidate technique is a literature review. This will reveal the relevant aspects of the group of domains the particular domain belongs to. If the system is simple, then these aspects could be translatable from the abstract domain to the specific domain, and it is less likely that relevant aspects of the simple domain are missed.

Even though a soft-inclusive approach is widely considered important for predictive modeling (Cao, Yu, Zhang, & Zhao, 2010; Gu & Tang, 2005; Shmueli, 2010; Shmueli & Koppius, 2011), there are few clear, rigorous approaches provided in literature. Often, domain understanding is an implicit part of model development. There is therefore a need for transparent, repeatable techniques that capture relevant understanding from complex domains.

3. Domain analysis in the supply chain

To find how domain analysis is used for simple and complex domains, we have performed a literature review. For our literature review, we focus on papers about predictive modeling for the supply chain. We have chosen this focus, as the supply chain is a well known complex field (New, 1997). Therefore, it gives a good overview of the domain analysis methods that are used in practice. The methodology and details of the results can be found at **XXX**.

3.1. Review Results

As mentioned in the previous section, we ranked each paper's domain complexity. Papers that had a complexity score > 3.00 have a *complex* domain, papers with a score ≤ 3.00 have *simple* domains. Papers that used soft analysis for variable selection or data cleaning are soft-inclusive, the others are soft-exclusive.

The majority of papers (96 % of the total number in the review) have a complex domain. These papers deal with real-world case studies in the supply chain The two papers that deal with a simple domain made predictions about a system made up of only intelligent agents.

84% (42/50) of the total papers and 87.5% (42/48) of the papers dealing with complex domains use a soft-exclusive domain analysis. Mostly, these papers use literature review as a means of variable selection. Domain analysis is generally not explicitly mentioned at all, and many papers use data analysis as a substitute for domain analysis. This means that variables are selected based on the provided data. Data cleaning is mentioned in only seven of the papers.

Most papers dealing with complex & soft-exclusive analysis are examples of Operations Research. The methodology of these papers tends to be to view cases as instances of well-known problems, resulting in a domain analysis based on literature about the well-known problem.

The six papers that did include soft factors mostly used interviews as a means of capturing domain intelligence. Rabelo *et al.* used Forrester's (2008) system dynamics, using both existing theory and interviews with those involved (Rabelo, Helal, Lertpattarapong, Moraga, & Sarmiento, 2008).

4. Requirements for a prediction instrument development method

The theoretical background makes it clear that soft-inclusive domain analysis is more appropriate for complex domains. It would appear that there is a considerable misalignment between the type of analysis that should theoretically be used for complex domains, and that which is used in practice.

The gap that should be addressed is how this analysis should be performed. Therefore, we present requirements below for a transparent prediction model development method with soft-inclusive domain analysis.

Design science approach The goal of the method is to design and develop an artefact: the *prediction instrument*. The field of design science research provides general methodologies on how to address such a goal (Gregor & Hevner, 2013; Peffers, Tuunanen, Rothenberger, & Chatterjee, 2008). The prediction instrument development (PID) method should follow these good practices, and consist of the same basic steps: *problem identification & objective definition*, *design & development*, *demonstration & evaluation*, and *communication* (Peffers et al., 2008).

Quantitative and qualitative methods The domain analysis part of the method should capture soft factors, which fits with the use of qualitative methods such as interviewing. In contrast, the development of predictive models uses quantitative methods such as statistical analysis and machine learning. Therefore, the combination of these two parts i.e. the PID method, is an example of *mixed methods* (Venkatesh, Brown, & Bala, 2013). In PID, the quantitative methods are used to validate the qualitative methods: predictive models are created to test those factors from the domain analysis that truly affect the prediction outcome.

Domain-driven For complex systems, the development of prediction instruments should start with domain analysis. Next, amongst other uses, the findings from the domain analysis are used for variable selection and data cleaning, which are subsequently used for predictive model development.

There are six types of intelligence that should be gathered in the domain analysis part of PID: data intelligence, network intelligence, social intelligence, human intelligence, organization intelligence, and domain intelligence (Cao et al., 2010). These are synthesized into four constraints: a *data constraint*, which states the quantity and quality of available data, and therefore rules out predictive models that require more or better data; an *interestingness constraint*, which states the minimum performance needed from the model; a *deployment constraint*, which determines what infrastructure predictive models have to be able to interact with; and a *domain constraint*, which contains laws & business rules the predictive model should be in accordance with, but also determines what performance measure to use, and any other reason why a predictive model would not be actionable.

Builds on existing methodologies This is not a stringent requirement, but given that there is a multitude of methodologies for predictive model development, such as that by Shmueli & Koppius (Shmueli, 2010; Shmueli & Koppius, 2011), and more generically for data mining, the Cross Industry Standard Process for Data Mining (Chapman et al., 2000). Using parts of an existing method, ensures greater validity and generalizability. These methodologies represent best practices for predictive model development, and should therefore be included in the PID method.

Efficiency As a final requirement, the PID method should be efficient. As it comes in place of implicit domain analysis or other opaque methods of gathering domain intelligence, it should provide as little burden as possible. This means that no unnecessary predictive models should be developed.

4.1. Intelligence meta-synthesis for Prediction Instrument Development

In this section, we describe a method for the development of a prediction instrument that captures and uses soft factors from the domain. Our method is based on intelligence meta-synthesis, which has previously been used to create economic models (Gu & Tang, 2005) and as an analysis tool for association rule mining (Cao et al., 2010). Below, we describe how we have adapted the method for the purpose of PID.

The method consists of three stages. In stage I, *qualitative assumptions* are gathered from the domain. For PID, these would be assumptions or hypotheses on factors influencing what is predicted. The stage starts with a field study, brainstorm or interview with domain experts to determine all relevant hypotheses. This is called *hypothesis divergence*. The unique and interesting hypotheses are then determined in the *hypothesis convergence* step. In stage I, the domain experts are also consulted to determine the domain constraints.

The whole of stage II is based on Shmueli & Koppius' method for predictive analytics (Shmueli, 2010; Shmueli & Koppius, 2011), and consists of variable selection, data cleaning, algorithm selection, and evaluation. In stage II, predictive models are developed based on the domain analysis from stage I. The hypotheses are used as a basis for variable selection and data cleaning. The data & domain constraints are used to determine which hypotheses can be tested (a hypothesis can only be tested if there is enough data of sufficient quality available). Next, using selected algorithms, predictive models are created. These are then evaluated, to determine their predictive performance relative to the interestingness constraint.

In stage III, the predictive models from stage II are validated qualitatively. The models that have sufficient performance and that do not violate the domain or deployment constraints are presented to the decision makers from the domain. Then, either by discussion, voting, or a technique like the Analytic Hierarchy Process (Saaty, 2008), one predictive model is selected. If the decision makers decide that no predictive model is good enough for whatever reason, or if all the predictive models were eliminated by the constraints, parts of the IMS-PID process can be repeated. If there a predictive model is selected, that forms the basis for the prediction instrument.

5. Discussion

Based on the theoretical background discussed at the start of this paper (see Section 2), it is important to include soft factors when analysing a complex domain for the purposes of prediction instrument development. The key aspect is not only to understand the domain, but also to understand how it is represented/filtered by the data. If the domain and the filter between domain and data are not well understood, manipulated data, missing variables, or other types of noise in the data could remain undiscovered, potentially severely limiting predictive performance. This is especially the case in complex domains, where human behavior, politics, myths & meanings and other soft factors play a role.

The results of our literature review show that although the vast majority of papers deal with such a complex domain, only a small minority explicitly mention the use of soft-inclusive domain analysis. Domain analysis in any form, such as a literature review, is only mentioned in thirty of the 50 papers. In this section, we discuss this apparent mismatch between theory and actual research practice. We consider three scenario's for the little or no mention of soft-inclusive domain analysis in the papers from the literature review: the researchers had pre-existing expertise on the domain they are working on, soft-inclusive domain analysis was performed implicitly, or domain analysis was not performed at all.

If the researchers developing a prediction instrument have sufficient understanding of the domain they are working on, due to past expertise, they could find performing a detailed domain analysis unnecessary. Whether the researchers' expertise on a domain is sufficient, is difficult to quantify, but here the 'proof is in the pudding', i.e. if the predictive models have adequate predictive performance, that is proof enough of the researchers understanding the domain. As the supply chain & transport domains are well researched, it is likely that many researchers have enough experience to quickly understand new domains.

So, only if researchers believe that they do not have enough expertise with the specific domain they are working on, or if the predictive performance is inadequate, a domain analysis would be performed. As so few papers in the review mention the use of domain analysis (even including literature review), we hypothesize that *Domain analysis is not performed* in these cases.

6. Conclusion

In this paper, we have investigated the role of domain analysis in prediction instrument development. In particular, we have investigated the relation between the type of domain (complex or simple) and the type of domain analysis (soft-inclusive or soft-exclusive). Below, we present our conclusions.

In Section 2, we presented arguments for the use of domain analysis for prediction instrument development. The data used to create prediction instruments is a filtered representation of the domain: not all aspects of the domain are translated to data, and furthermore, not all aspects are translated directly. It is important that both the domain and filter are well understood. If not, important variables could be missing in the data, the data could contain more noise than anticipated, and data manipulations could remain undiscovered.

Especially for complex systems, an analysis based on similar domains is not a good fit. The soft factors within the domain, such as human behavior, make it incomparable to an abstraction of that domain (Checkland, 1989; Forrester, 1994). Therefore it is important for domain analysis to be soft-inclusive, i.e. to include soft factors that are specific to that domain.

To understand how domain analysis is conducted in practice, we conducted a structured literature review of the use of predictive analytics in the field of transport and the supply chain. As this field is inherently complex due its many interconnected parts and strong human component (New, 1997), it gives a good view of how domain analysis is performed in practice. The review resulted in a total of fifty papers

Ninety-six percent (48/50) of the papers from the review deal with a complex domain. Only 12.5 percent (6/48) of those papers explicitly discuss the use of soft-inclusive domain analysis. As these papers do show sufficient predictive power, it is likely that the researchers have in fact acquired the necessary understanding of the domain.

To make the process of understanding a complex domain more transparent, explicit, and rigorous, we have collected requirements for a soft-inclusive domain analysis method that needs to be a part of prediction instrument development. Based on these requirements, we have developed a method based on a combination of Gu and Tang's *intelligence meta-synthesis* and Shmueli & Koppius' method for predictive analytics (Gu & Tang, 2005; Shmueli, 2010; Shmueli & Koppius, 2011). Our method, which we call *intelligence meta-synthesis for prediction instrument development* or IMS-PID, uses brainstorming and a field study as a method for collecting domain knowledge. This domain knowledge comes in the form of hypotheses on how to make predictions and constraints that determine which predictive models are usable & interesting. Next, these hypotheses are translated to a strategy for data cleaning & variable selection, which forms the basis for predictive model development. The constraints determine which models can be pursued further, and which can be abandoned, e.g. due to lack of available data, non-conformity with business rules, or a lack of predictive power.

In summary, the primary contributions of this paper are as follows: (i) we have shown the theoretical importance of soft-inclusive domain analysis for the purpose of prediction instrument development, (ii) we have shown that in practice, papers are mostly implicit on how domain analysis is done, and finally (iii) to address this issue, we have developed IMS-PID: a transparent, explicit and rigorous method for prediction instrument development that includes soft-inclusive domain analysis.

References

Association for Information Systems. (2014). Senior Scholars' Basket of Journals. Retrieved from http://aisnet.org/?SeniorScholarBasket

Cao, L., Yu, P. S., Zhang, C., & Zhao, Y. (2010). *Domain driven data mining* (1st). Springer Publishing Company, Incorporated.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Rüdiger, W. (2000). CRISP-DM 1.0.

Checkland, P. B. (1989). Soft systems methodology. *Human systems management*, *8*(4), 273–289.

Forrester, J. W. (1994). System dynamics, systems thinking, and soft or. *System Dynamics Review*, *10*(2–3), 245–256.

Fry, T. D. & Donohue, J. M. (2013). Outlets for operations management research: a dea assessment of journal quality and rankings. *International Journal of Production Research*, *51*(23-24), 7501–7526.

Gregor, S. & Hevner, A. R. (2013). Positioning and presenting design science research for maximum impact. *MIS Quarterly*, *37*(2), 337–355.

Gu, J. & Tang, X. (2005). Meta-synthesis approach to complex system modeling. *European Journal of Operational Research*, *166*(3), 597–614.

Hensher, D. (2011). Institute of Transport and Logistics Studies' journal rankings for transport, logistics and supply chain management. Retrieved from http://rstrail.tudelft.nl/sites/default/files/journals%7B%5C_%7Dtransport-logistics-journal-rankings-110530%7B%5C_%7D3.pdf

Jackson, M. & Keys, P. (1984). Towards a system of systems methodologies. *The Journal of the Operations Research Society*, *35*(6), 473–486.

Menachof, D. A., Gibson, B. J., Hanna, J. B., & Whiteing, A. E. (2009). An analysis of the value of supply chain management periodicals. *International Journal of Physical Distribution & Logistics Management*, *39*(2), 145–165.

Meredith, J. R., Steward, M. D., & Lewis, B. R. (2011). Knowledge dissemination in operations management: published perceptions versus academic reality. *Omega*, *39*(4), 435–446.

Mylonopoulos, N. A. & Theoharakis, V. (2001). On site: global perceptions of is journals. *Communications of the ACM*, *44*(9), 29–33.

New, S. J. (1997). The scope of supply chain management research. *Supply Chain Management: An International Journal*, *2*(1), 15–22.

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2008). A design science research methodology for information systems research. *Journal of Management Information Systems*, *24*(3), 45–77.

Rabelo, L., Helal, M., Lertpattarapong, C., Moraga, R., & Sarmiento, A. (2008). Using system dynamics, neural nets, and eigenvalues to analyse supply chain behaviour. a case study. *International Journal of Production Research*, *46*(1), 51–71.

Saaty, T. L. (2008). Decision making with the analytic hierarchy process. *Int. J. Services Sciences*, *1*(1), 83–98.

SCImago Journal & Country Rank. (2014). Journal Rankings. Retrieved from http://www.scimagojr.com/journalrank.php

Serenko, A. & Dohan, M. (2011). Comparing the expert survey and citation impact journal ranking methods: example from the field of Artificial Intelligence. *Journal of Informetrics*, *5*(4), 629–648.

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, 289–310.

Shmueli, G. & Koppius, O. R. (2011). Predictive analytics in information systems research. *MIS Quarterly*, *35*(3), 553–572.

Stonebraker, J. S., Gil, E., Kirkwood, C. W., & Handfield, R. B. (2012). Impact factor as a metric to assess journals where OM research is published. *Journal of Operations Management*, *30*(1), 24–43.

Thomson Reuters. (2014). Journal Citation Reports. Retrieved from http://thomsonreuters.com/journal-citation-reports/

van der Spoel, S., Amrit, C., & van Hillegersberg, J. (2015). Predictive analytics for truck arrival time estimation: a field study at a European distribution center. *International Journal of Production Research*, *forthcoming*.

van der Spoel, S., van Keulen, M., & Amrit, C. (2013). Process prediction in noisy data sets: a case study in a Dutch hospital. In *Data-driven process discovery and analysis* (pp. 60–83). Springer.

Venkatesh, V., Brown, S. A., & Bala, H. (2013). Bridging the qualitative-quantitative divide: guidelines for conducting mixed methods research in information systems. *MIS Quarterly*, *37*(1), 21–54.

Webster, J. & Watson, R. T. (2002). Analyzing the past to prepare for the future: writing a literature review. *MIS Quarterly*, *26*(2), 3.