A Visual Analytics Environment for Developing Data Quality-aware Performance Models

Marco Angelini$^{1[0000-0001-9051-6972]}$ and Cinzia Daraio$^{2[0000-0002-4825-0071]}$

Sapienza University of Rome
{angelini,daraio}@diag.uniroma1.it

Abstract. This paper proposes a Visual Analytics Environment to carry out a data quality-aware development of performance models on data that integrate heterogeneous sources looking for patterns of research performance.

Keywords: Visual Analytics · Data Quality · Performance Models · Ontology-based data integration.

1 Introduction

In the last decade, the rapid increase in the production, communication and evaluation of research have been signs of a transformation. Despite the various innovations introduced with big data, machine learning and altmetrics, the role of the user of metrics, and her interactions in the development and evaluation phase of performance models have not received the great attention they deserve. In addition, important aspects for the usability of data and information, such as the different dimensions of data and information quality, have frequently been overlooked, making the developed performance measurement systems rigid, fragile and inconsistent. In order to mitigate these problems, Daraio et al. [1] developed a data quality approach featured on higher educational data that are integrated with research data and other heterogeneous sources through Sapientia, the Ontology of Multidimensional Research Assessment. In Daraio et al. [2] there is a description of the benefits of an ontology-based data integration approach for data quality in an open environment. In [3] we showed the advantages of Visual Analytics for the development of performance models, i.e. models capable to describe the behaviour of subjects with respect to an analysis goal. A performance model can be an efficiency model (e.g. Free Disposal Hull) that describes how efficiently Universities manage their scientific publications production with respect to other factors (e.g. public and private funding, staff, etc.). In this paper we make a step further and extend the flexibility of a visual analytic approach featured to performance models development to include data quality procedures and tests.

Few previous works exist that have explored the use of Visual Analytics to conduct data quality analysis, in particular for supporting research activities evaluation. Among the most relevant, Liu et al. [4] proposed a literature review on Visual Analytics for data quality activities, and a framework for conducting data cleansing on four data types (multimedia, text, trajectories and graphs). Gschwandtner et al. [5] proposed a solution for data cleansing of time-oriented data, providing semi-automatic quality checks, visualizations, and directly editable data tables. However the authors specifically target time-series and do not specifically target research activities evaluation. We highlight as a differentiating point with respect to these approaches that this proposal is expressively aimed at research evaluation activities, taking into account the specific indicators and semantic that govern this domain. At the same time, this proposal shares similar goal in allowing identification of data quality supporting the performance models and steering of this quality toward the desired level.

2 Approach

This paper exploits Visual Analytics, "the science of analytical reasoning facilitated by interactive visual interfaces" [6] focusing on the data quality analysis of the measures, indicators, scores that will be used by the analyst as a base for creating a performance model. In this respect this phase is very important, given the heterogeneity of data sources, the different formats that can still convey similar semantic, and the importance that features selection can have on the definition of a performance model. The authors contributed in [3] a workflow for dynamically creating and assessing the quality of a performance model for evaluating research activities. This workflow is based on an ontological modelling of the data sources, instantiated in the *Sapientia* ontology, that align semantically the contents coming from different sources (e.g. Eter, Scopus). From this step, a Visual Analytics Environment is built that allows us to explore the data and build on top of them several performance models (e.g. Efficiency models, input/output models) that can be compared and assessed in order to be validated. Figure 1 shows the mentioned functionalities as steps 1 and 3.

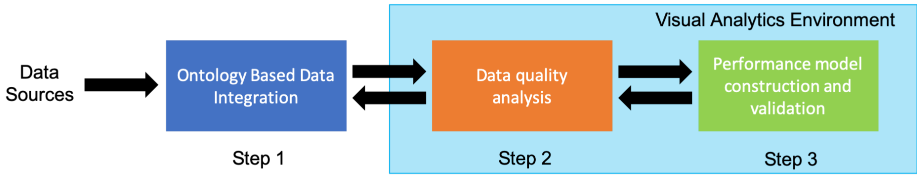

Visual Analytics Environment
Data Sources
Ontology Based Data Integration
Data quality analysis
Performance model construction and validation
Step 1
Step 2
Step 3

Fig. 1. Workflow for the construction and validation of Performance models.

The proposal in this paper leverages on this workflow inserting a new intermediate step (see Figure 1 step 2) that implements the evaluation of the quality of data ingested in the system, during the data exploration and/or once the analyst selected a pool of features on which construct the desired models (hence the bi-directional arrows for both model construction and data ingestion). For quality we mean both syntactic properties of the data, like the presence of null or incomplete values or the type of data at hand (e.g. categorical, numerical, etc..), and semantic properties, like the fairness of specific features (consistency), their timeliness and their comparability. This intermediate step can reinforce the resulting quality of the developed performance models, and helping in better respecting some characteristics like fairness of the model or control the reliability of the obtained performance rankings with respect to the statistical significance of the supporting data. Given the specificity of the task, the resulting Visual Analytics Environment has been expanded with a tailored dashboard dedicated to this analysis, constructed with both visual paradigms that are familiar to data quality experts (e.g. Pareto charts) and more abstract representations, like matrix based visualization or radar-charts. Focusing on the second class, an example is visible in Figure 2: this representation, inspired by the work of Angelini et al. [7], allows to statistically describe the behaviour of several data dimensions with respect to a set of data quality metrics.

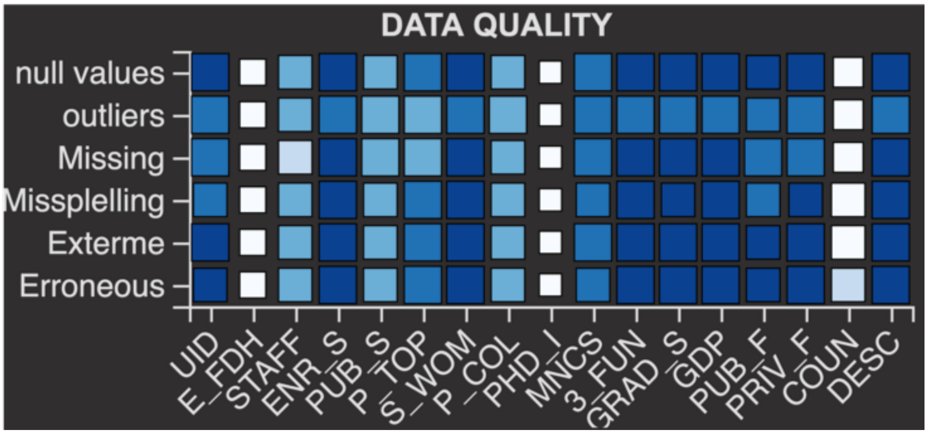
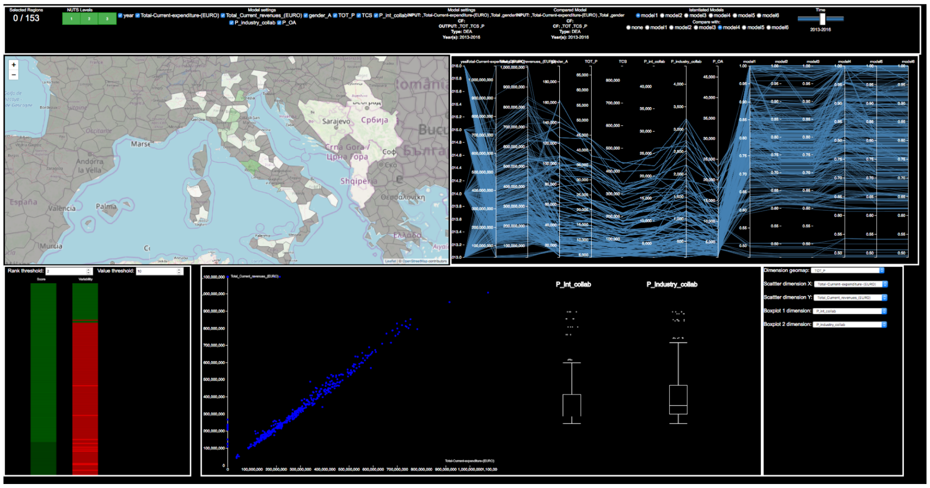


Fig. 2. Data dimensions quality view. This matrix-like visualization represents data dimensions on x-axis and quality metrics on y-axis. The visualization provides an overview of data quality for the whole dataset, highlighting problematic dimensions or quality checks.

This matrix-like visualization represents data dimensions on x-axis and quality metrics on y-axis. It visually highlights good quality scores (deep blue color) or bad quality scores (white color) for data dimensions on all quality metrics. This results in the creation of an an overview on the overall quality of the dataset through a compact visualization that allows exploration and detailed analysis of subset of pairs quality metric(s)-data dimension(s). The area of an element encodes the variability in scores computed for all the data tuples for each pair quality metric-data dimension. A large area means a narrow confidence interval (stable results, low variability). Conversely, a small area identifies a large confidence interval and an high variability in scores.



Selected Regions 0 / 153

NUTS Levels 1 2 3

☑ year ☑ Total-Current-expenditure-(EURO) ☑ Total_Current_revenues_(EURO) ☑ gender_A ☑ TOT_P ☑ TCS ☑ P_int_collab ☑ P_industry_collab ☑ P_OA

Model settings
INPUT: ,Total-Current-expenditure-(EURO) ,Total ,gender
CF:
OUTPUT: ,TOT ,TCS ,P
Type: DEA
Year(s): 2013-2016

Compared Model
INPUT: ,Total-Current-expenditure-(EURO) ,Total ,gender
CF:
CF: ,TOT ,TCS ,P
Type: DEA
Year(s): 2013-2016

Istantiated Models
● model1 ○ model2 ○ model3 ○ model4 ○ model5 ○ model6

Compare with:
○ none ○ model1 ○ model2 ○ model3 ● model4 ○ model5 ○ model6

Time 2013-2016

Dimension geomap: TOT_P
Scattter dimension X: Total-Current-expenditure-(EURO)
Scattter dimension Y: Total_Current_revenues_(EURO)
Boxplot 1 dimension: P_int_collab
Boxplot 2 dimension: P_industry_collab

Fig. 3. A view of the Visual Analytics Environment that supports the creation and validation of data quality-aware research performance models.

Combining these visual paradigms, the resulting environment, implementing the described workflow, is visible in Figure 3 for the performance model evaluation part. All its visualizations are empowered by powerful interaction means for governing the identification of not satisfactory quality level and potential improvements. This environment has been used and validated during a Methodological Course organized within the training activities of the EU RISIS Project (Research Infrastructure for Science and Innovation Policy Studies, www.risis2.eu). Further development of the work is on-going, and a novel version integrating received comments and more comprehensive functionalities will be integrated in an online platform available for future evaluation activities.

References

1. C. Daraio, R. Bruni, G. Catalano, A. Daraio, G. Matteucci, M. Scannapieco, D. Wagner-Schuster, and B. Lepori, "A tailor-made data quality approach for higher educational data," *Journal of Data and Information Science*, vol. 5, no. 3, pp. 129 – 160, 2020.
2. C. Daraio, M. Lenzerini, C. Leporelli, P. Naggar, A. Bonaccorsi, and A. Bartolucci, "The advantages of an ontology-based data management approach: Openness, interoperability and data quality," *Scientometrics*, vol. 108, p. 441–455, July 2016.
3. M. Angelini, C. Daraio, M. Lenzerini, F. Leotta, G. Santucci, *et al.*, "Performance model's development: A novel approach encompassing ontology-based data access and visual analytics.," in *Scientometrics*, p. to appear, 2020.

4. S. Liu, G. Andrienko, Y. Wu, N. Cao, L. Jiang, C. Shi, Y.-S. Wang, and S. Hong, "Steering data quality with visual analytics: The complexity challenge," *Visual Informatics*, vol. 2, no. 4, pp. 191 – 197, 2018.
5. T. Gschwandtner, W. Aigner, S. Miksch, J. Gärtner, S. Kriglstein, M. Pohl, and N. Suchy, "Timecleanser: A visual analytics approach for data cleansing of time-oriented data," in *Proceedings of the 14th International Conference on Knowledge Technologies and Data-Driven Business*, i-KNOW '14, (New York, NY, USA), Association for Computing Machinery, 2014.
6. K. A. Cook and J. J. Thomas, "Illuminating the path: The research and development agenda for visual analytics," tech. rep., Pacific Northwest National Lab.(PNNL), Richland, WA (United States), 2005.
7. M. Angelini, V. Fazzini, N. Ferro, G. Santucci, and G. Silvello, "Claire: A combinatorial visual analytics system for information retrieval evaluation," *Information Processing & Management*, vol. 54, no. 6, pp. 1077 – 1100, 2018.