Exploratory Causal Analysis with Time Series Data

Synthesis Lectures on Data Mining and Knowledge Discovery

Synthesis Lectures on Data Mining and Knowledge Discovery is edited by Jiawei Han, Lise Getoor, Wei Wang, Johannes Gehrke, and Robert Grossman. The series publishes 50- to 150-page publications on topics pertaining to data mining, web mining, text mining, and knowledge discovery, including tutorials and case studies. Potential topics include: data mining algorithms, innovative data mining applications, data mining systems, mining text, web and semi-structured data, high performance and parallel/distributed data mining, data mining standards, data mining and knowledge discovery framework and process, data mining foundations, mining data streams and sensor data, mining multi-media data, mining social networks and graph data, mining spatial and temporal data, pre-processing and post-processing in data mining, robust and scalable statistical methods, security, privacy, and adversarial data mining, visual data mining, visual analytics, and data visualization.

Exploratory Causal Analysis with Time Series Data
James M. McCracken
2016

Mining Human Mobility in Location-Based Social Networks
Huiji Gao and Huan Liu
2015

Mining Latent Entity Structures
Chi Wang and Jiawei Han
2015

Probabilistic Approaches to Recommendations
Nicola Barbieri, Giuseppe Manco, and Ettore Ritacco
2014

Outlier Detection for Temporal Data
Manish Gupta, Jing Gao, Charu Aggarwal, and Jiawei Han
2014

Provenance Data in Social Media
Geoffrey Barbier, Zhuo Feng, Pritam Gundecha, and Huan Liu
2013

Graph Mining: Laws, Tools, and Case Studies
D. Chakrabarti and C. Faloutsos
2012

Mining Heterogeneous Information Networks: Principles and Methodologies
Yizhou Sun and Jiawei Han
2012

Privacy in Social Networks
Elena Zheleva, Evimaria Terzi, and Lise Getoor
2012

Community Detection and Mining in Social Media
Lei Tang and Huan Liu
2010

Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predictions
Giovanni Seni and John F. Elder
2010

Modeling and Data Mining in Blogosphere
Nitin Agarwal and Huan Liu
2009

Exploratory Causal Analysis with Time Series Data
James M. McCracken

ISBN: 978-3-031-00781-1 paperback
ISBN: 978-3-031-01909-8 ebook

DOI 10.1007/978-3-031-01909-8

A Publication in the Springer series
SYNTHESIS LECTURES ON DATA MINING AND KNOWLEDGE DISCOVERY

Lecture #12
Series Editors: Jiawei Han, *University of Illinois at Urbana-Champaign*
Lise Getoor, *University of Maryland*
Wei Wang, *University of North Carolina, Chapel Hill*
Johannes Gehrke, *Cornell University*
Robert Grossman, *University of Chicago*
Series ISSN
Print 2151-0067 Electronic 2151-0075

Exploratory Causal Analysis with Time Series Data

James M. McCracken
George Mason University

SYNTHESIS LECTURES ON DATA MINING AND KNOWLEDGE DISCOVERY #12

ABSTRACT

Many scientific disciplines rely on observational data of systems for which it is difficult (or impossible) to implement controlled experiments. Data analysis techniques are required for identifying causal information and relationships directly from such observational data. This need has led to the development of many different time series causality approaches and tools including transfer entropy, convergent cross-mapping (CCM), and Granger causality statistics.

A practicing analyst can explore the literature to find many proposals for identifying drivers and causal connections in time series data sets. Exploratory causal analysis (ECA) provides a framework for exploring potential causal structures in time series data sets and is characterized by a myopic goal to determine which data series from a given set of series might be seen as the primary driver. In this work, ECA is used on several synthetic and empirical data sets, and it is found that all of the tested time series causality tools agree with each other (and intuitive notions of causality) for many simple systems but can provide conflicting causal inferences for more complicated systems. It is proposed that such disagreements between different time series causality tools during ECA might provide deeper insight into the data than could be found otherwise.

KEYWORDS

time series causality, leaning, exploratory causal analysis

To Angel, whose patience I have tested often but never broken

Contents

Preface

Consider a scientist wishing to find the driving relationships among a collection of time series data. The scientist probably has a particular problem in mind, e.g., comparing the potential driving effects of different space weather parameters, but a quick search of the data analysis literature would reveal that this problem is found in many different fields. They would find proposals for different approaches, most of which are justified with philosophical arguments about definitions of causality and are only applicable to specific types of data. However, the scientist would not find any consensus of which tools consistently provide intuitive causal inferences for specific types of systems. The literature seems to lack straightforward guidance for drawing causal inferences from time series data. Many of the proposed approaches are tested on a small number of data sets, usually generated from complex dynamics, and most authors do not discuss how their techniques might be used as part of a general causal analysis.

This work was developed from the realization that drawing causal inferences from time series data is subtle. The study of causality in data sets has a long history, so the first step is to develop a loose taxonomy of the field to help frame the specific types of approaches an analyst may be seeking (e.g., time series causality). Then, the philosophical causality studies must be carefully and deliberately divorced from the data causality studies, which is done here with the introduction of *exploratory causal analysis* (ECA). Finally, examples need to be presented where the different approaches are compared on identical data sets that have strongly intuitive driving relationships. Using such an approach, the analyst can develop an understanding of how a causal analysis might be performed, and how the results of that analysis can be interpreted. This work presents all three of these steps and is intended as an introduction and guide to such analysis.

James M. McCracken
March 2016

Acknowledgments

This work would not be have been possible without the help of many people who have supported me over the last few years, financially and otherwise. I have, hopefully, made those people aware of how much I appreciate their help. Most importantly though, I'd like to thank my adviser Professor Robert Weigel. I appreciate that he was willing to let me prove that I was a serious student, despite my somewhat unusual circumstances as a doctoral candidate. His academically holistic advising style emphasized a deep understanding of both the background literature and the many different methods for approaching modern data analysis problems. I believe he has helped me learn not only the material printed in this lecture but, more importantly, how to be a professional physicist. I could not have had a better adviser.

James M. McCracken
March 2016