Contents

Methods **1**

Results & Discussion **1**
Time line 1
Chord diagram 1
Normality 3
Maps 3
Network analysis 3
LDA performance 3

Methods

Our two-fold, novel and comprehensive research approach combines advanced computation with a stakeholder survey to describe past water research in Latin America. First, we performed a data-driven literature review by assembling a corpus of 30,000 water resources research articles and analyzing them with a topic model. We used Latent Dirichlet Allocation (LDA, [@Blei2003]), a generative Bayesian model, which describes topics as a probability distribution over words and documents as a probability distribution over topics. Human reading validated the document topics and identified the country of study of 2,000 articles. Combined with article metadata and text mining, this information was used to predict the country of study across the corpus with machine learning. In-corpus citing and cited references were used to build a citation network which, combined with topic and location information, infers connectivity between research communities.

Second, to understand the landscape of water research in Latin America, we collected publicly available data and conducted an on line survey. Countries within Latin America were statistically clustered into four groups with distinct physiographic and socioeconomic characteristics. To ground our data-driven results in the reality of the current research climate, we invited nearly 20,000 corresponding authors to share their experiences through a survey focused on research discipline, accessibility and connectivity.

Results & Discussion

Time line

The scientific production of water research based in Latin America and the Caribbean has grown exponentially over the last four decades, revealing a significant contribution of knowledge that this region is providing to the global scientific community. We grouped research by country cluster, which we calculated using 43(?) variables, to identify how countries with similar social and hydrological systems contributed to water research over time. The growth in research is not distributed equally across the region and is dominated by contributions from the two largest countries in Latin America, Brazil and Mexico. A residual analysis of the exponential growth of annual research by country cluster identifies trends in annual scientific output. Since 2000, the first year when we have over 30 research papers in each language, there were three distinct periods of water research. Annual output was lower than the general trend for the first several years of the 21st century, followed by a period of relatively higher output and ends with a trend of decreasing growth. There may be a connection between Brazil's economic crisis in 2012….

Chord diagram

A chord diagram describes the composition of water research in Latin America and the Caribbean and reveals inequalities in locations and themes of research (figure 1). The chord widths indicate the proportion of a specific research theme within the top 25% of research for a given country. While Brazil, Mexico, Argentina and Chile dominate the research landscape, countries in the Caribbean and most of Central America are excluded from the analysis due to their statistically insufficient number of articles (less than 30), indicating a relative shortage of research in these regions. A country's socio-economic cluster correlates to its contribution

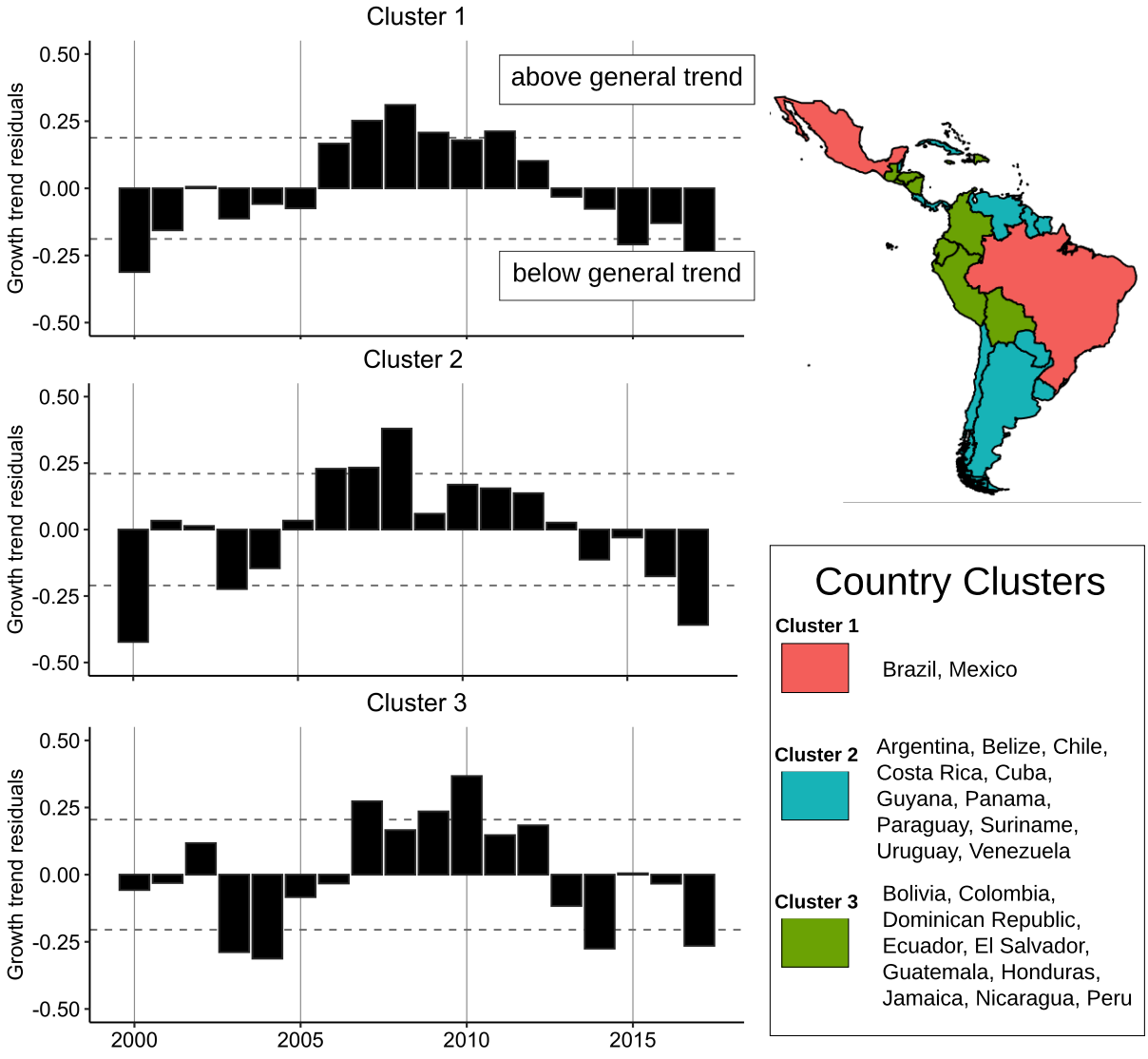

Figure 1: Time line

to overall research output, suggesting that a country's resources, geography and history influence the scientific activity of researchers working there. Similarly, water research is not distributed equally among disciplines and there is a relative shortage of research in the social sciences. While Mexico contributes most to the social science research, it is a small proportion of its overall output. Water research is conducted primarily in the physical and life sciences, with Mexico and Argentina alternating for second highest output after Brazil, respectively.

After assessing trends in the corpus, we further analyzed results from the topic model and text mining to identify bright spots and blind spots of water research in Latin America and the Caribbean. We define successful research as having a distribution that is close to the standard normal distribution and with high entropy. When applied to our corpus, these concepts highlight areas within water resources research that are relatively under-studied.

Normality

Water research in Latin America and the Caribbean has generally higher normality across countries than documents (figure 2). A review of the normalities of the components of the water budget validates our analysis approach. Rivers and precipitation, which must be monitored and understood to manage water resources, have distributions closest to normal, while glaciers are far from normal distribution because few countries have glaciers to study. Assuming that high normality indicates success, we identify statistics, quantitative methods and water sampling as bright spot of research methods. Niche topics, such as irrigation and isotopes, have high normality across countries but low normality across documents and lie somewhere between bright and blind spots. Such research topics are either infrequently mentioned or, if mentioned, are the main subject of a paper and lack integration in interdisciplinary research. The least normality is seen in two topics of great importance for water management: reservoirs and risk assessment.

Maps

Network analysis

LDA performance

All of this research

1/3 of the NSF specific categories present in our corpus are present in all 3 languages

Importantly, most of the top 10% of research is captured in thee categories in all 3 languages (9 of 13)

Another 1/3 is in 2 languages and the last 1/3 is only present in English

Results from the topic model can accurately describe the research occuring in all 3 languages in most categories