Action

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.

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 the Caribbean and most of Central American are excluded from the analysis, indicating a relative shortage of research in these regions. A country's socio-economic cluster correlates to its contribution to overall research output, which suggests that a country's resources, geography and history influence its scientific activity. Similarly, water research is not distributed equally among disciplines and results from our corpus reveal 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 distributions that is normally and with high entropy. When applied to our corpus, these concepts highlight locations and themes that are relatively under-researched.

Water research in Latin America and the Caribbean has generally higher normality across countries than documents (figure 2). A look at 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 normally distribution because few countries have glaciers to study. This analysis clearly identifies the main research methods that are used in water research, which include spatial methods, statistics, quantitative methods and water sampling. Following the decreasing gradient of normality points next to the niche topics, such as irrigation and isotopes, which have high normality across countries but low normality across documents. These values indicate that topics in this region of the graph are not often discussed with other research topics or included in interdisciplinary papers. The least normality is seen in two topics of great importance for water management: reservoirs and risk assessment.