



HLB Research and
Extension Outputs Network

Decoding HLB research: Using text mining to support grower decision-making

Ulianova Vidal-Gomez¹, Kaity Coleman², Sandra Olkowski², Neil McRoberts², David McClure³, Uros Mojsilovic³, Mike Hill³, Dharmendra Kalauni⁴, Sebastian Galindo⁴, Lauren M. Diepenbrock¹, Tripti Vashisth¹, Yu Wang¹, and Megan M. Dewdney¹

¹Citrus Research and Education Center, University of Florida, Lake Alfred, FL. ²University of California, Davis, Davis, CA. ³Purdue University, West Lafayette, IN., ⁴University of Florida, Gainesville, FL.

INTRODUCTION

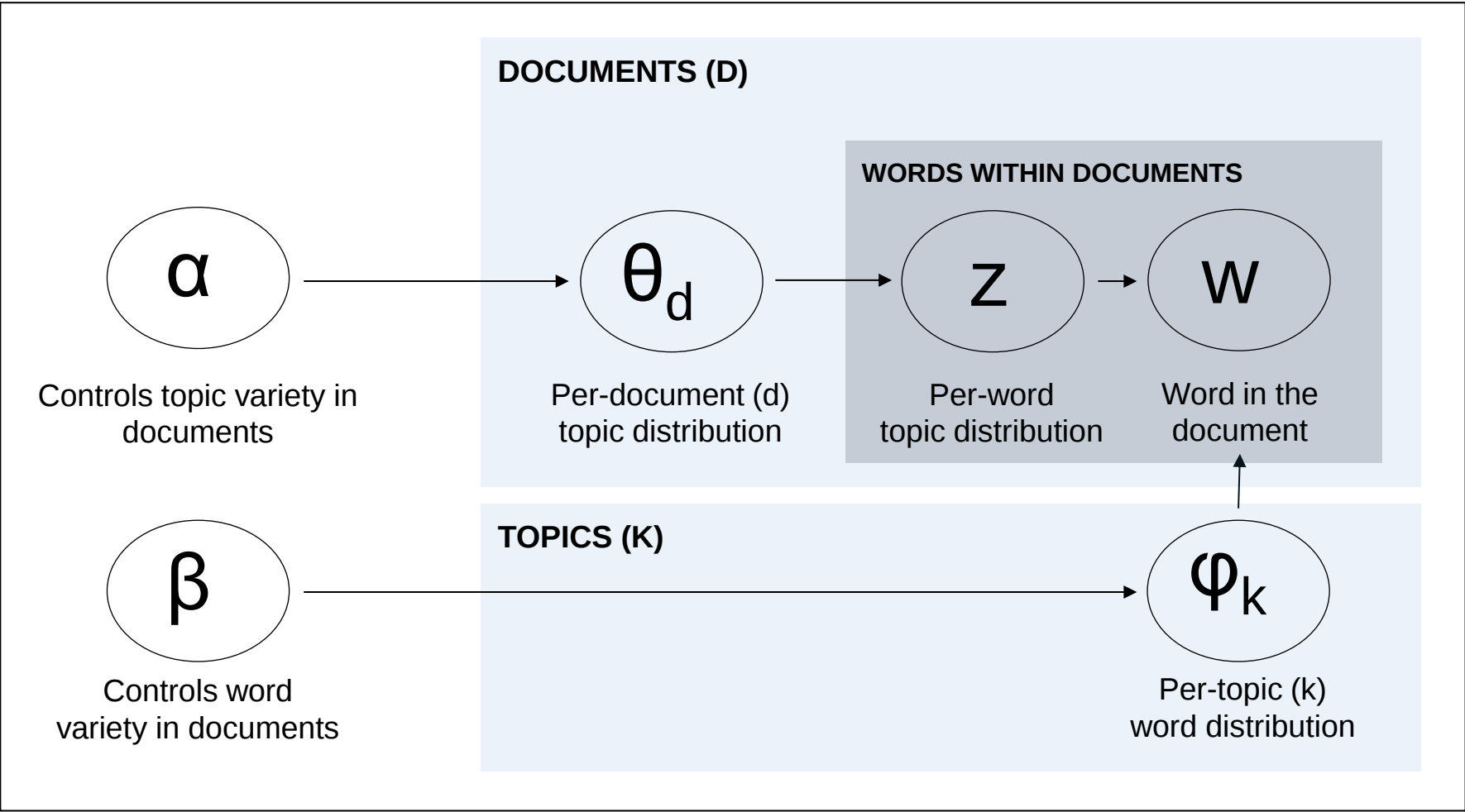
Over the past decade, Huanglongbing (HLB) has drastically reduced citrus yields in the United States. In response, the USDA’s National Institute of Food and Agriculture (NIFA) has funded hundreds of research projects aiming to understand and manage HLB. However, despite this substantial investment, the resulting knowledge remains fragmented, highly technical, and difficult for end users—particularly growers—to access or apply. To synthesize and evaluate this body of work, we analyzed 269 NIFA-funded HLB grants, using text mining techniques to identify dominant research themes, underexplored areas, and opportunities to align future efforts with more immediate, field-level needs.

*United States Department of Agriculture (USDA)

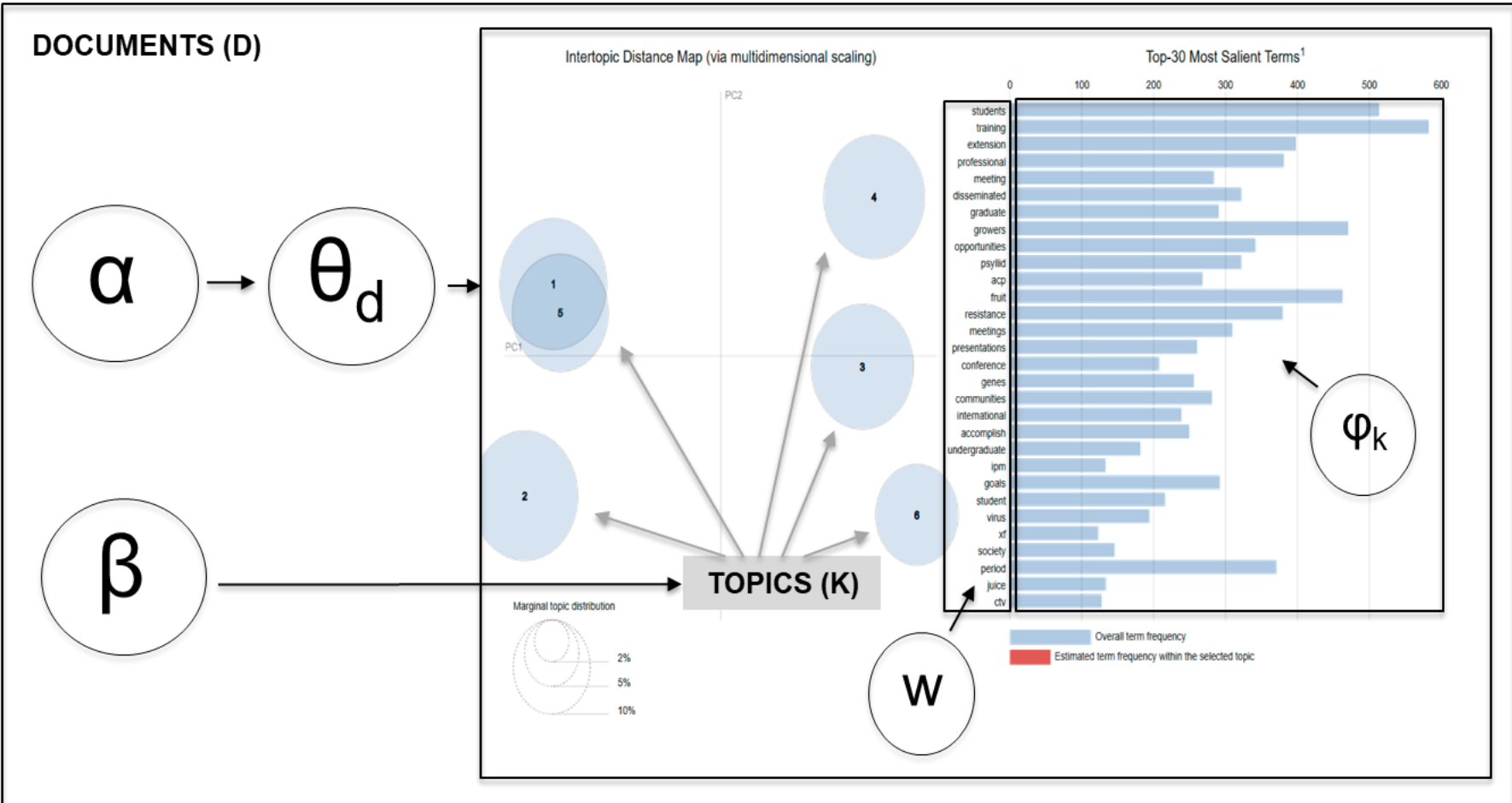
METHODS

- Analyzed 254 USDA-funded HLB grants from the **USDA NIFA Project Reporting Portal (Gateway)**.
- Used each project’s non-technical summary and outputs (872 text entries)
- From preprocessing, a document-term matrix was created with >12,000 unique terms. We applied:
- **Word frequency, co-occurrence, and clustering** to explore key terms and relationships
- **Latent Dirichlet Allocation (LDA)** to identify six hidden research themes based on word distributions
- **Principal Component Analysis (PCA)** and **UMAP** to visualize topic patterns across grants
- Topics interpreted using top 100 terms (β)
- Dominant topics assigned to each document based on posterior probabilities (θ).

Latent Dirichlet Allocation (LDA) Model:



LDA Outputs:



Analyses were conducted in R 4.4.1, using “tm”, “tidytext”, “topicmodels”, “tidyverse”, “textclean”, “forcats”, “umap”, “LDAvis”, “servr”, “writextl” and “readr” libraries (R core team 2024).

RESULTS

SECTION 1: Textual similarity clustering of NIFA HLB grants summaries and outputs (UMAP)

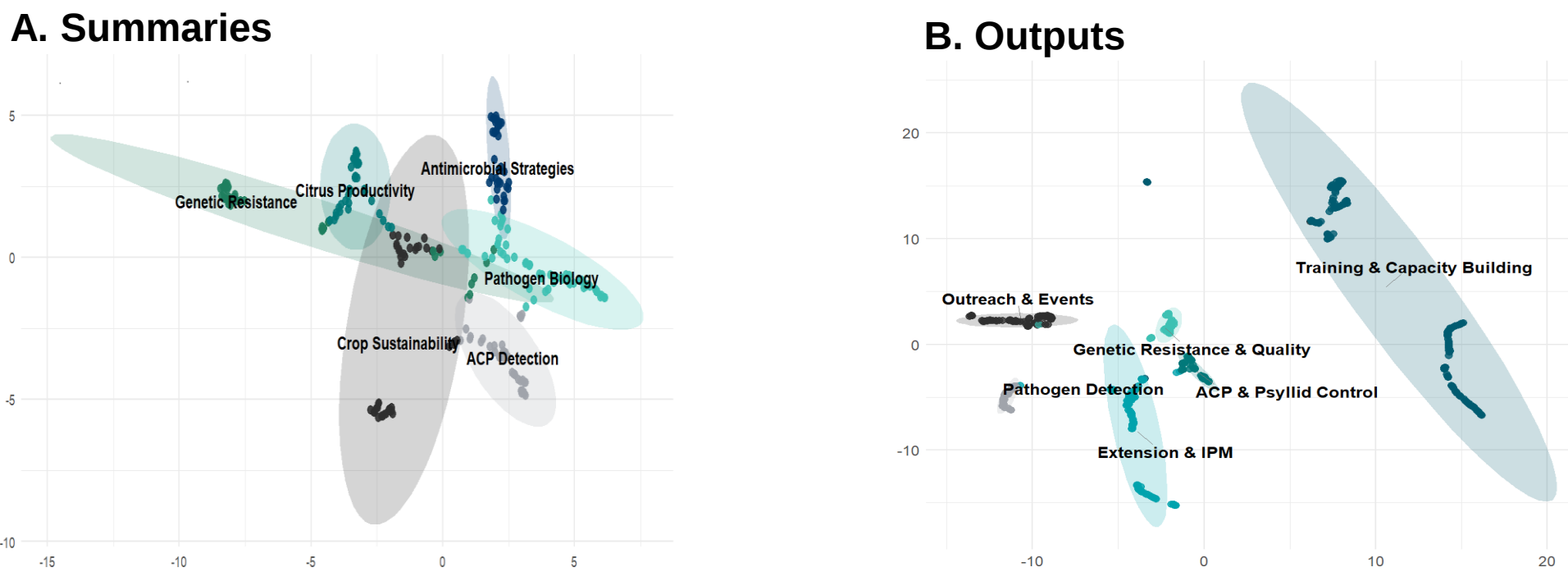


Fig. 1. UMAP projection showing the similarity among 254 NIFA-funded HLB research grants based on textual content from A. Summaries and B. Outputs. Each point represents a grant. Grants with similar topic compositions appear closer together in the 2D space, revealing thematic clusters across the corpus.

SECTION 2. Topic modeling of NIFA HLB grant summaries (LDA analysis)

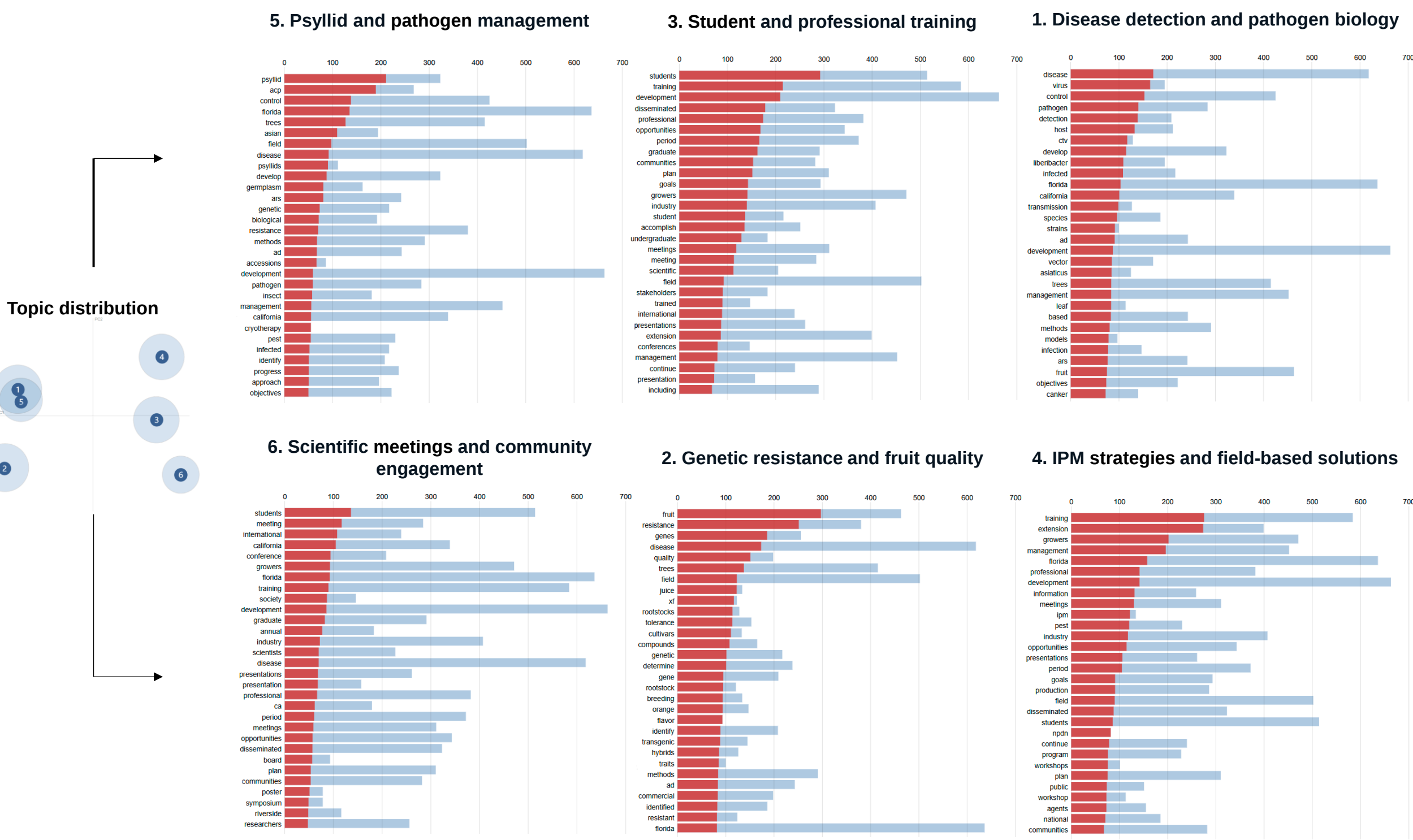


Fig. 2. LDA identified six topics across NIFA-funded HLB research outputs. Each set of outputs is represented as a distribution of these topics, with a dominant theme assigned based on θ . The topic distribution diagram shows how topics are positioned in relation to each other based on term similarity. Bar graphs display the top 30 terms per topic: red bars highlight the most distinctive terms, while gray bars represent additional contributing terms. Topic numbers are assigned arbitrarily by the algorithm and do not reflect any order.

DISCUSSION

1. Topics in summaries focus on training, outreach, and capacity building—reflecting institutional goals at the proposal stage. Outputs, meanwhile, emphasize applied areas like antimicrobial treatments and nutrition, signaling an effort toward extension and capacity building, but not necessarily the immediate management strategies needed in the field to address HLB’s impact.
2. Vector management is a consistent theme in both, reflecting strong and sustained research investment.
3. Summaries emphasize NIFA priorities, while outputs reflect more field-oriented goals typical of other USDA agencies.
4. The impact of capacity-building efforts remains unclear and unquantified.
5. LDA helped reveal thematic shifts, guiding actions for researchers, policymakers, and growers.

CONCLUSION

Understanding how research priorities shift from proposals to outputs reveals critical gaps and opportunities. While vector control remains a consistent focus, more actionable strategies and measurable impact—especially in capacity building—are needed to strengthen HLB management across the citrus industry.

REFERENCES

- USDA (n.d.). *NIFA Reporting Portal*. U.S. Department of Agriculture. Retrieved from <https://portal.nifa.usda.gov>
- We used Latent Dirichlet Allocation (LDA; Blei et al., 2003) for topic modeling, visualized topics with LDAvis (Sievert & Shirley, 2015) following principles by Chuang et al. (2012), and applied UMAP (McInnes et al., 2018) for 2D projection.