THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

MSc DATA SCIENCE

ST498 – Capstone Project proposal

Project Title: Estimation of missing tariff and trade data

Thomas Verbeet ¦ World Trade Organization ¦ Thomas.Verbeet@wto.org

Project description

Tariff and trade data are essential for analyzing market access, global trade policies, and economic trends. The World Trade Organization's (WTO) Integrated Database shows some data gaps, creating significant obstacles to accurate economic modelling and policy analysis. Addressing these gaps is critical to achieving a comprehensive and reliable dataset.

This capstone project aims to develop and implement robust methodologies to estimate missing tariff and import data. The methodologies developed will apply to future datasets, improving data coverage and accuracy. This project will utilize publicly available and non-public datasets, including notifications to the WTO, ITC MacMap and UN Comtrade, and other relevant tariff and trade data repositories. The focus will be on creating a systematic, replicable approach to data imputation and validation, enhancing the usability of trade data for research, policy, and economic analysis.

The main objectives/milestones of this project are:

1. **Familiarize yourself with Existing Datasets**

- Review the WTO Integrated Database, for both tariff and trade data.
- Understand their structure, coverage, and areas where data gaps exist.

2. **Undertake Literature Research**

- Conduct a comprehensive review of relevant literature on handling missing data in large-scale datasets, with a focus on trade and economic data.
- Examine the latest statistical and machine learning techniques for data imputation and validation.

3. **Identify and Develop Methodologies for Estimating Missing Data**

- Explore and define statistical and machine learning-based methods for estimating missing values and address the issue of preferential tariff misreporting.

4. **Develop an R/Python Program for Data Imputation**

- Create a Python or R-based interpolation algorithm to implement the selected methodologies, automating the imputation of missing tariff and trade data.
- Ensure scalability and ease of use for future data updates.

- Implement baseline methods for comparison, such as mean imputation, forward/backward filling and others.

5. **Model evaluation & comparison**

- Evaluate imputation accuracy using common metrics (e.g. RMSE, MAE, R^2), cross-validation, comparing methods on a holdout dataset with artificially induced missing values or others

6. **Create a Reusable Framework and Write a Verification & Data Quality Module**

- Develop a systematic framework that can be applied to similar datasets.
- Build a verification module that ensures data quality through cross-validation with secondary sources and implements sensitivity analysis and quality control mechanisms to flag inconsistencies.

7. **Write a Working Paper or Background Note on the Methodology**

- Document the research findings, methodologies used, and the results of the data imputation process.
- Prepare a background note that details the application of the developed framework and its potential for future use in trade data analysis.

Expected Impact:

The development of a reliable methodology for filling data gaps will significantly improve the accuracy of tariff and trade datasets, supporting more informed policy decisions and economic research. By creating a reusable framework and data quality verification tool, the project will contribute to the long-term improvement of available tariff and trade data for analysts, economists, and policymakers at large.

References

WTO Tariff Analysis Online
UN Comtrade
Market Access Map (macmap.org)

WTO, Modalities and Operation of the Integrated Database, G/MA/367, 2019 (in particular Annex 5 on tariff recomposition)
https://docs.wto.org/dol2festaff/Pages/SS/directdoc.aspx?filename=q:/G/MA/367.pdf&Open=True

Teti, Feodora, "Missing Tariffs, False Imputation, and the Trade Elasticity", 2023
https://www.econstor.eu/bitstream/10419/277636/1/vfs-2023-pid-86676.pdf