NEW YORK ECONOMIC REVIEW



JOURNAL OF THE NEW YORK STATE ECONOMICS ASSOCIATION

cc BY NC ND

VOLUME LVI

NEW YORK STATE ECONOMICS ASSOCIATION

FOUNDED 1948

2025-2026 OFFICERS

President:

KPOTI KITISSOU • SUNY ONEONTA

Vice-President:

OLUWASHEYI OLADIPO • SUNY OLD WESTBURY

Secretary:

MICHAEL McAVOY • SUNY ONEONTA

Treasurer:

PHILIP SIRIANNI • SUNY ONEONTA

Director of communications:

KASIA PLATT • SUNY OLD WESTBURY

Editorial Board, *New York Economic Review*:

ARINDAM MANDAL • EDITOR • SIENA UNIVERSITY
CYNTHIA BANSAK • ASSOCIATE EDITOR • ST. LAWRENCE UNIVERSITY
ANIRUDDHA MITRA • ASSOCIATE EDITOR • BARD COLLEGE
JOSEPH ONOCHIE • ASSOCIATE EDITOR • BARUCH COLLEGE
KASIA PLATT • ASSOCIATE EDITOR • SUNY, OLD WESTBURY
BEN J. NIU • ASSOCIATE EDITOR • ST. JOHN FISHER UNIVERSITY

Conference Proceedings Editor

MANIMOY PAUL • SIENA UNIVERSITY

Board of Directors:

L. CHUKWUDI IKWUEZE • QUEENSBOROUGH COMMUNITY COLLEGE, CUNY
LAUREN CALIMERIS KOCMAN • ST. JOHN FISHER UNIVERSITY
BEN J. NIU • ST. JOHN FISHER UNIVERSITY
LEV VLASENKO • SUNY POTSDAM
BABITA SRIVASTAVA • WILLIAM PATTERSON UNIVERSITY
CLAIR SMITH • ST. JOHN FISHER COLLEGE
JANHAVI SHANKAR TRIPATHI • ST. BONAVENTURE UNIVERSITY
CHRISTINE STORRIE • SUNY ONEONTA
YIZHAO (ANDY) WANG, SUNY CANTON

The views and opinions expressed in the Journal are those of the individual authors and do not necessarily represent those of individual staff members or the New York State Economics Association.

Vol. 56, Fall 2026

CONTENTS

ARTICLES

An Applied Use of Exploratory Spatial Data Analysis in the Identification of Poverty Clusters in the Buffalo- Niagara Falls, New York Metropolitan Statistical Area: 2008-2022

Craig Rogers……..……..5

Diversification in Electricity Price and Volume Swaps

Matthew Brigida, David Brigida & Kathleen Maceyka.……..……20

Comparative Analysis of Inflation Forecasting using Machine Learning: Evidence from Selected G20 Countries

Rolando A. Santos & Brian Sloboda ……..……..……..…....30

Causal Diagrams and Structural Debriefing in Macroeconomics: A Literacy-Targeted Pilot Study

Elena Smirnova, Natalia Smirnova & Oleg Pavlov……..……49

Referees..……..……..……72

EDITORIAL

The *New York Economic Review* (NYER) is an annual publication of the New York State Economics Association (NYSEA). The NYER publishes theoretical and empirical articles, and also interpretive reviews of the literature in the fields of economics and finance. All well-written, original manuscripts are welcome for consideration at the NYER. We also encourage the submission of short articles and replication studies. Special Issue proposals are welcome and require a minimum of 4 papers to be included in the proposal, as well as a list of suggested referees.

MANUSCRIPT GUIDELINES

All manuscripts are to be submitted via https://www.nyseconomicsassociation.org/new-york-economic-review/submit-paper-to-the-editor/ to the editor in a Microsoft Word file format, with content formatted according to the guidelines at:

https://www.nyseconomicsassociation.org/NYER/Guidelines

The NYER is cited in:

- EconLit by the American Economic Association
- Ulrich's International Periodicals Directory
- Cabell's Directory of Publishing Opportunities in Business and Economics
- Australian Business Deans Council (ABDC) Journal Quality List

ISSN NO: 1090-5693

An Applied Use of Exploratory Spatial Data Analysis in the Identification of Poverty Clusters in the Buffalo- Niagara Falls, New York Metropolitan Statistical Area: 2008-2022

Craig Rogers*

ABSTRACT

Poverty research in the 1970s and the 1980s focused on metropolitan areas and major central cities of the largest metropolitan areas. Since the 1990s, poverty research literature has examined poverty within the context of urban and rural areas; counterbalancing political ideologies and the changing political landscape and their resultant poverty policies.

This paper takes a different approach using the new American Community Survey (ACS) 5-year datasets and exploratory spatial data analysis within a geospatial framework to analyze the changing intra-metropolitan location of poverty in the Buffalo-Niagara Falls, NY metropolitan statistical area between 2008-2022.

INTRODUCTION

Sixty years after the federal government's declaration of a War on Poverty, the United States still has a fragmented public policy response to poverty. Unquestionably, poverty levels in the U.S. did decrease during the height of the federal response in the late 1960s and early 1970s. The progress against poverty has been uneven since and there exist persistent spatial pockets of poverty in U.S. metropolitan areas. The uneven progress against poverty is due partially to the limitation of the headcount index as a measure of poverty. Poverty measurement is a complex undertaking that involves analysis of demographic and socioeconomic characteristics within a spatial environment.

This paper differs from other metropolitan focused poverty research initiatives in that: (1) uses a consistent census tract based spatial data set to track the changes in the location of poverty; (2) delineates the metropolitan area into cities, inner and outer ring suburban areas; and (3) uses exploratory spatial data analysis (ESDA) within a GIS environment to explore the spatial distribution and patterns of intra-metropolitan poverty.

POVERTY MEASUREMENT LITERATURE REVIEW

Poverty measurement is an important indicator of a nation's economic and social status. Poverty measurement affects not only the perception of well-being; it affects numerous federal, state, and local public policy formation, program administration, and public understanding (Citro and Michael, 1995). The

* Department of Economics and Finance, 2001 Main Street, Canisius University, Buffalo, NY 14208.

most widely used poverty measure in the United States is known as the official poverty rate and is constructed as a headcount index derived originally from the work of Orshanky (1965).

The official headcount poverty measure is the most used poverty measure in the United States, and this paper uses this measure with the acknowledgment of its empirical and interpretative limitations. Criticism of the headcount poverty measure identifies three limitations: (1) the measure does not take into consideration the intensity of poverty; (2) the measure does not measure the depth of poverty; and (3) the measure is computed based on households not individuals. Headcount index criticism is well founded and has led to the misrepresentation of the poverty status in the United States and has underestimated the reduction in the overall poverty level in the U.S. (Primus, 2006). The Panel on Poverty and Family Assistance (Citro and Michael, 1995) suggests the headcount index weakness has become more apparent and consequential because of far-reaching changes in the U.S. society, national and global economies and in government policies.

Poverty researchers suggest one single measure of poverty; fails to capture the complexity of the underlying poverty issues and additional measures of poverty should be used as a means of understanding the depth and severity of poverty which may reveal important differences in the quantitative and qualitative nature of poverty (Osberg and Xu, 1999). These differences would suggest different types of poverty reduction initiatives for different spatial areas, Jolliffe (2003).

There has been significant advancement in the theory and methods of poverty measurement since Sen's (1976) article on poverty measurement. Unfortunately, Myles and Picot (2000) acknowledge poverty measurement advancement has been slow to move beyond academic poverty researchers and has not been embraced by policymakers. They note that a major reason for this is that these measurements have at their core abstract mathematical properties and do not provide policymakers or the public with an easily understood poverty measure.

In 2005, the U.S. Census Bureau introduced the American Community Survey (ACS). The ACS is an ongoing survey that provides information on the demographic, economic, and social economic characteristics of the nation at different spatial levels for either 1-year or 5-year intervals. The ACS discontinued the 3-year ACS due to budgetary issues in 2015. Poverty data at the census tract level is only available using the 5-year ACS.

Poverty research literature is extensive. Sawhill (1988) offers one of the more comprehensive summaries of U.S. poverty and within the context of specific time periods provide a summary of the evolution of the poverty literature, the main highlights and observations, and general research patterns.

Poverty research in the 1970s and the 1980s focused on metropolitan areas and major central cities of the largest metropolitan areas. Since the 1990s, poverty research literature has examined poverty within the context of urban and rural areas; counterbalancing political ideologies and the changing political landscape and their resultant poverty policies (Cushing and Zheng, 2000). The writings of Wilson in the late 1980's focusing on the underclass led to a renewal in poverty research, particularly at a micro spatial level. (Wilson and Aponte 1985).

There has been little research on the changes in the distribution of poverty at the intra-metropolitan level and more specifically in which the intra-metropolitan area has been defined by central cities, inner and outer ring suburban areas (Johnson, 2007). Urban researchers recognize the limitation of using an elementary dichotomy of central city and non-central city suburban areas in public policy analysis. Dual conceptualization of urban areas into cities and non-center (suburban) areas does not recognize the significant difference in suburban areas due to different demographic, socioeconomic, and political characteristics (Kneebone, 2009).

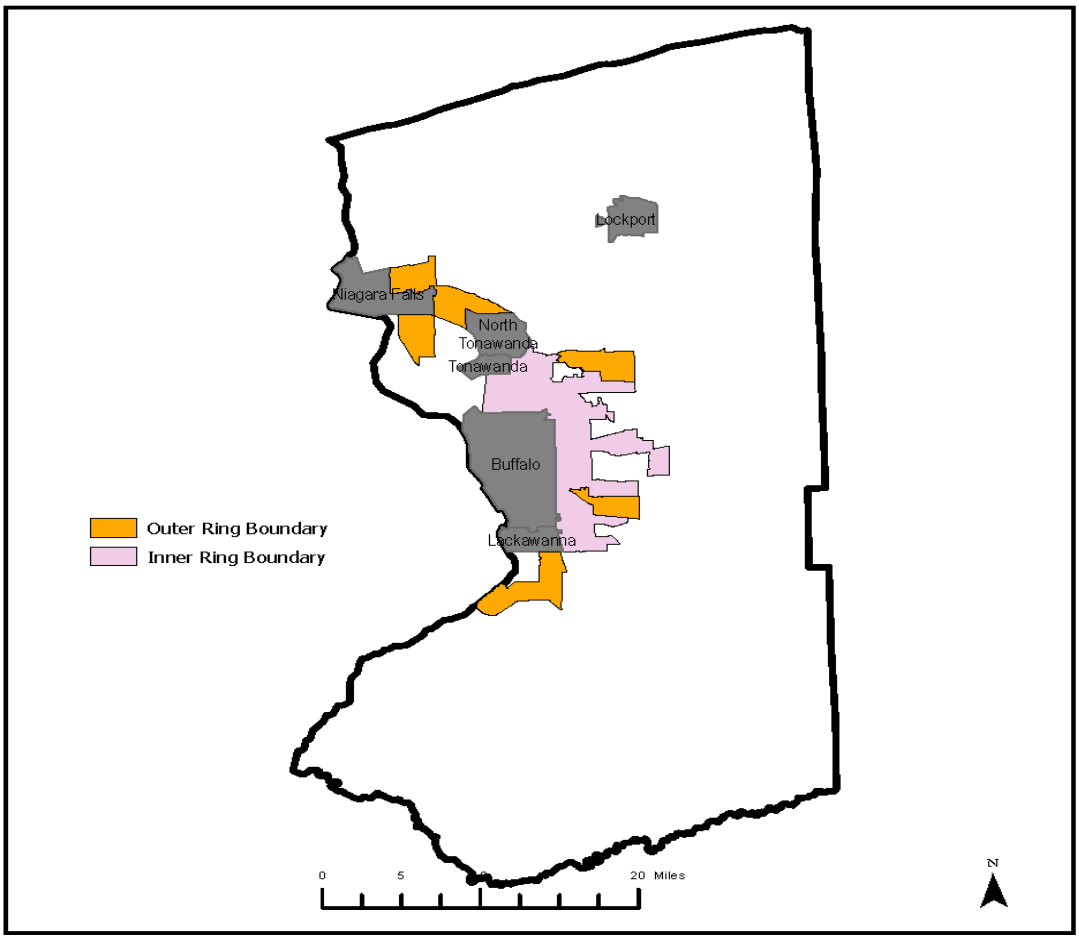
One spatial approach used to analyze change at the intra-metropolitan level is to construct spatial boundaries using census tracts. A census tract is a spatial boundary that has between 1200 – 8000 people with an average population of 4000 people. A census tract boundary is usually a contiguous area but may vary by spatial size; central city census tracts are spatially smaller than census tracts in rural areas.

Spatial boundaries for cities and the suburban areas in the Buffalo-Niagara Falls metropolitan statistical area (MSA) were generated from census tract-based boundary shape files from the U.S. Census Bureau. Inner and outer ring suburban areas were constructed using a modified version of a research method suggested by Leigh and Lee (2005). Using their census tract-based typology, inner ring suburban census tracts comprised non-central city census tracts that had a population per square mile density of 1000 people and 400 housing units per square mile built between 1950 and 1970. The outer ring suburban census tracts comprised non-central city and non-inner ring census tracts that had 200 housing units per square mile built between 1970 and 2000 and met the population per square mile requirement of 1000 people. Map 1 shows the location of central cities, and inner and outer-ring suburban areas in the Buffalo MSA. The Buffalo-Niagara Falls, New York MSA consists of Erie and Niagara counties and six central cities: Buffalo, Niagara Falls, Lackawanna, Tonawanda, North Tonawanda, and Lockport.[1]

A recent thread of neighborhood poverty research examines poverty, its spatial concentration and its short and long-term socioeconomic impact of residents living in high poverty neighborhoods. The long-term consequences of concentrated high poverty neighborhoods in cities have begun to receive attention across interrelated academic disciplines such as sociology, political science, public health, economics, and geography (Bhatta, 2001; and Bellu and Liberati, 2005). Census tracts have been used extensively in poverty research examining poverty neighborhood effects (Cooke and Marchant 2006). Neighborhood poverty research using census tracts as the spatial boundary have used two different poverty rates to identify high poverty census tracts: 30% (Cooke and Marchant, 2006); and 40% (Strait, 2001).

Studies have shown there exists neighborhood effects on the short and long-term impact on education, employment, and criminal behavior linked to living in high poverty neighborhood (Chyn and Katz, 2021). Studies examining the relocation of residents from high poverty neighborhoods to low poverty neighborhoods have shown different long term socioeconomic outcomes based on age, gender, and

Map 1. Cities, Inner and Outer Ring Suburban Areas in Buffalo MSA



duration of living in high poverty neighborhoods. (Chetty and Hendren 2018; and Chetty, Hendren, and Katz, 2016). The residential relocation of a young person from a high poverty neighborhood to a low poverty neighborhood improved their educational and employment outcomes; the relocation of an adult from a high poverty neighborhood to a low poverty neighborhood did not have significant impact on their educational and employment outcomes.

The association between neighborhood poverty and individual current and long-term outcomes may be due to two different mechanisms (Chyn and Katz, 2021). Neighborhood environments can have a causal impact on adults and children; the other possibility there exists a non-random sorting of people who live in high poverty census tracts (Oreopoulos, 2003; and Laliberte, 2021).

A myriad of demographic, socioeconomic, and political factors have been identified which contribute to poorer central city neighborhoods with a high percentage of racial and ethnic minority residents increasing their economic and social isolation. The research of Wilson (1987), Kasarda (1990), Massey and Denton (1993) and others have identified and explored how deindustrialization, urban sprawl, the decline of blue-

collar manufacturing jobs, limited public transportation, and suburbanization of entry level employment opportunities, and poorer quality inner city educational systems have increasingly led to the concentration of minority residents in poverty-stricken neighborhoods.

EXPLORATORY SPATIAL DATA ANALYIS WITHIN A GEOSPATIAL ENVIRONMENT

Geospatial technologies are defined by three inter-related systems and tools that acquire, analyze, manage, store, or visualize spatial data. The geospatial technology field is interdisciplinary and consists of three areas: (1) geographic information systems (GIS); (2) remote sensing; and (3) global positioning systems (GPS). Geographical Information Systems (GIS) has emerged as a mapping and spatial analysis tool that is widely used in education, government, and industry to identify and solve spatial issues and problems.

Geographical information systems (GIS) is one of the fastest growing segments of the computer industry, ranging across disciplines from the physical sciences to city planning Pittman and Thrall (1991). Managing the economic and social development of a nation, metropolitan area, city, or suburban area is a complex and challenging endeavor increasingly relying on the integration of spatial statistics and GIS to assist policymakers at all spatial levels to efficiently use scarce resources and develop greater accountability in measurable outcomes (Li, Henry, & Davis, 1999).

GIS use in many organizations primarily revolves around basic map making and does not tap into the full capacity of GIS to move beyond this basic function into a broader use of the technology in examining patterns of economic, demographic, and socioeconomic change taking place at different spatial levels in metropolitan areas. There has been a steady improvement in the inclusion of spatial statistics in GIS software. Users of GIS-based spatial analysis tools must work across standalone programs, and this limits the functionality and use of spatial statistics in organizations. A contributor to the limited use of exploratory spatial data analysis (ESDA) methods within a GIS environment is the limited exposure of staffing to spatial statistical methods. Boots and Okabe (2007) suggest that non-GIS specialists may find it more difficult to determine the most appropriate tool for their analysis given the increased expansion of GIS based tools for spatial analysis.

This paper focuses on the use and application of exploratory spatial data analysis (ESDA) methods in identifying the spatial distribution of poverty in the Buffalo metropolitan area. ESDA is defined as the collection of techniques to describe and visualize spatial distribution, identify atypical locations, discover patterns of spatial association, and suggest different spatial regimes and other forms of spatial instability or spatial non-stationarity as noted by Anselin (1993). ESDA should aim to: (1) describe spatial distributions; (2) discovery of spatial association patterns; (3) identification of different spatial regimes; and (4) finally, identification of spatial outliers (Anselin, 1993) and (Wilhelm & Steck, 1998).

Non-spatial statistical methods have three limitations that generated the need to develop alternative statistical tools for use in spatial analysis using spatial data. These three limitations are: (1) spatial data will show some degree of spatial dependence; (2) spatial data sets often are subjected to a modifiable areal

unit problem; and (3) the assumption of variable data stationarity are often violated across a study area and fueled the development of exploratory spatial data analysis methods. (Unwin & Unwin, 1998).

ESDA covers a wide spectrum of research methods and techniques; ESDA statistics are grouped into two classes: global and local. Global statistics are summary measures of an entire study region. These global measures can be grouped according to the attribute data used in the analysis, for interval or ratio scale data the following ESDA techniques are appropriate for Moran's *I* or Getis-Ord G analysis.

Wong and Lee (2005) provide an overview of the major characteristics of these Global Moran I is an inferential statistic that tests for global spatial autocorrelation in group level data; simultaneously using location and attribute values. Global Moran I analysis determines if the spatial pattern is clustered, dispersed, or random.

Results are interpreted in reference to a spatial null hypothesis. The null hypothesis is based on the presumption that the attribute analyzed is randomly distributed in the study area. There are two interpretations of a Global Moran I value: (1) if the p-value is not statistically significant, then you cannot reject the null hypothesis that the observed spatial distribution is random; and (2) if the p-value is statistically significant and the z-score is positive, then you may reject the null hypothesis, there is a greater probability the spatial distribution of values is clustered.

Like Global Moran *I*, Getis-Ord General G is an inferential statistical test interpreted within the context of a null hypothesis. The null hypothesis is no clustering of attribute values. A major difference in the interpretation of the z-score of General G test from the Global Moran *I*. There are three possible General G z-score interpretations: (1) the p-value is not statistically significant, and the null hypothesis is rejected (2) the p-value is statistically significant and the z-score is positive. In this case, the null hypothesis is rejected and there exists a clustering of "high" values in the dataset; and (3) the p-value is statistically significant, and the z-score is negative; in this case, you may again reject the null hypothesis and there exists a clustering of "low" values in the dataset.

Recent research in quantitative geography has placed a focus on identifying and understanding differences across spatial data rather than similarities. The emerging emphasis on local ESDA statistics is based on the idea that when analyzing spatial data, it might be incorrect to assume that the results obtained from the whole data set represent the situation in all parts of the study area (Fotheringham, Brunsdon, & Charlton, 2000). Wong and Lee (2005) note the limitations of global statistics in that a global value will mask *local* variation in the data variable in the study area. To capture the variance in spatial autocorrelation, local measures of spatial autocorrelation called *Local Indicators of Spatial Association (LISA)* are used.

The LISA family of indicators include "local" versions of Anselin's (1995) *Moran's I*, and Getis and Ord's (1992) *Gi* (G-statistic).* Local Moran's *I* developed by Anselin (1995) is a local measure of spatial autocorrelation in which spatial data are distributed so that high values are located near to other high values and low values are near other low values. The *G-statistic* measures the concentration of low or high data values; a large positive Z value indicates a concentration of large values close to one specified

neighborhood (i.e., *i)* and a large negative *Z* value implies many small values close to neighborhood *i.* This local statistical measure attempts to quantify local concentrations.

Two challenging elements of using LISA techniques are the identification of spatial neighborhoods and the computation of spatial weights. Spatial weights play a critical role in computation of spatial statistics, and there is no standard methodology used to identify and compute these weights. Spatial weights are frequently misspecified by the overreliance on spatial weights that are empirically convenient but have no obvious relationship to the structural dependence of the matrix (Getis & Aldstadt, 2004)**.**

There are multiple methods used to define neighborhood matrices for use in spatial data analysis.[2] The methods used are dependent on the type of spatial data used; for area data such as a census tract, adjacency is more appropriate than distance to define neighborhoods (Wilhelm and Steck, 1998). Because of the challenges in defining a spatial neighborhood and the limitations of using distance as a part of a spatial weight it is best to use two or more different spatial weight schemes to assess the robustness of the local spatial indicators used in the spatial analysis**.** This paper took this common approach and used two weighted matrices: (1) contiguity (edges only); and (2) inverse distance. The ESDA analysis in this paper is based on using a contiguity spatial weight matrix[3] . Unlike an inverse distance with its spatial weight decreasing as the distance between census tracts increases; a contiguity spatial weight defines neighbors by shared census tract boundaries. Wong and Lee (2005) suggest using different spatial autocorrelation statistics in spatial analysis to capture different aspects of spatial autocorrelation in the data.

This paper uses both global and local indicators of spatial analysis (LISA) as a means of developing a comprehensive understanding of the changing concentration and distribution of poverty in the Buffalo metropolitan area. This paper uses a commonly available commercial GIS software program to conduct exploratory spatial data analysis (ESDA), ESRI's ArcPro is a leading desktop platform GIS software program that has ESDA capabilities.[4]

EXPLORATORY SPATIAL ANALYSIS OF INTRA-METROPOLITAN POVERTY IN THE BUFFALO MSA

Poverty in the Buffalo MSA differs by intra-metropolitan location (Table 1). Historically, cities in the Buffalo metropolitan area experience the highest poverty rates; poverty rate in cities increased at a more rapid pace during the 1980s and 1990s and by year 2000 was almost double the poverty rate of the metropolitan area. There has been an increase in the poverty rate in inner and outer ring suburban areas in the Buffalo metro.

The American Community Survey (ACS) provides margin of error data that allow for constructing confidence intervals at the 90% confidence level (Field, 2009). Tables 1-1B highlights the intra-metropolitan poverty rate confidence intervals at four spatial scales: the Buffalo MSA; cities, inner ring, and outer rings areas in the Buffalo metropolitan area The lower bound confidence interval ranges are different across the four spatial levels; the cities' lower bound range is approximately four times the inner ring lower bound range and four to five times the outer ring range. The upper bound descriptive analysis of the intra-metropolitan poverty data highlights upper bound ranges that are different; the cities' upper bound values

are from twice to three times the inner and outer ring upper bound values. A two-step testing approach was used in the analysis.

Table 1. Buffalo MSA Intra-metropolitan Poverty Rate Confidence Intervals: 2008-2012

	Lower Bound	Mean	Upper Bound
Buffalo MSA	8.99%	13.99%	18.29%
Cities	18.86%	25.07%	31.17%
Inner Ring	4.64%	8.96%	12.74%
Outer Ring	4.23%	7.23%	9.94%

Table 1A. Buffalo MSA Intra-metropolitan Poverty Rate Confidence Intervals: 2013-2017

	Lower Bound	Mean	Upper Bound
Buffalo MSA	9.67%	14.62%	18.88%
Cities	19.81%	26.74%	32.32%
Inner Ring	5.70%	9.66%	13.09%
Outer Ring	3.92%	6.69%	9.19%

Table 1B. Buffalo MSA Intra-metropolitan Poverty Rate Confidence Intervals: 2018-2022

	Lower Bound	Mean	Upper Bound
Buffalo MSA	8.00%	13.61%	17.91%
Cities	16.49%	24.04%	29.42%
Inner Ring	4.32%	8.92%	12.49%
Outer Ring	4.29%	7.50%	10.09%

Source:U.S. Census Bureau. *Poverty Status in the Past 12 months, 2008-2012, 2013-2017, 2018-2022 American Community Survey 5-year estimates* [Table S1701] @ 90% confidence level

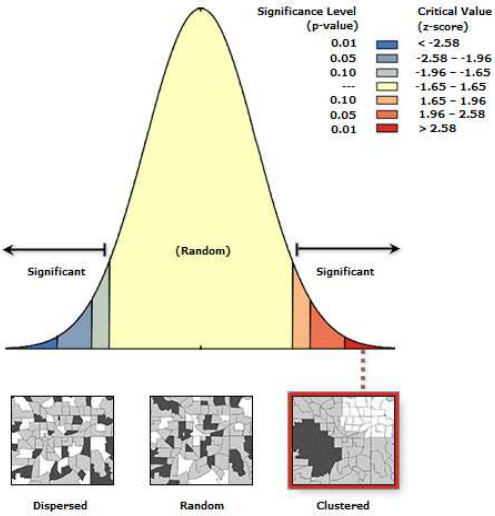
of the exploratory spatial data: (1) test for spatial autocorrelation using Global Moran's *I*; and (2) based on the global spatial autocorrelation result, test for local spatial autocorrelation.

Global Moran *I* analysis (Graph 1) shows within the Buffalo MSA there exists a non-random clustering of census tract poverty rates. Global Moran *I* index values for each study period had p-values that were statistically significant and their respective z-scores were positive leading to a rejection of the null hypothesis of a random distribution of census tract poverty rates in the Buffalo MSA[5].

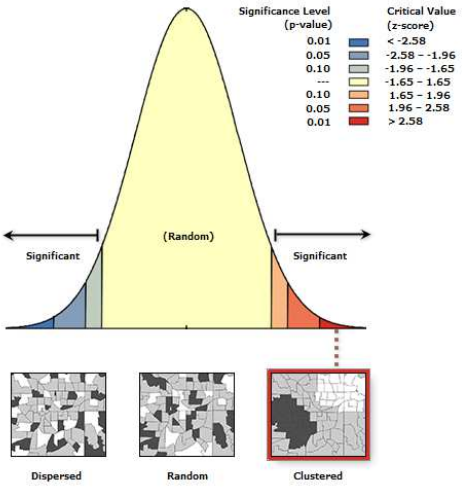
A limitation of Global Moran *I* analysis is its inability to spatial identify within the spatial boundary where spatial clustering exists. Local Moran *I* analysis is needed to identify within a spatial boundary the location of spatial clustering of spatial data. Local Moran *I* is the "local" version of Global Moran *I*; developed by Anselin (1995) measures spatial autocorrelation of spatial data identifying the location of high values.

Graph 1. Buffalo MSA Global Moran *I*: 2008 - 2022

Poverty Rate 2008-2012



Poverty Rate 2018-2022



located near other high values and/or low values. An important application of Local Moran *I* is the identification of spatial clusters or spatial outlier.

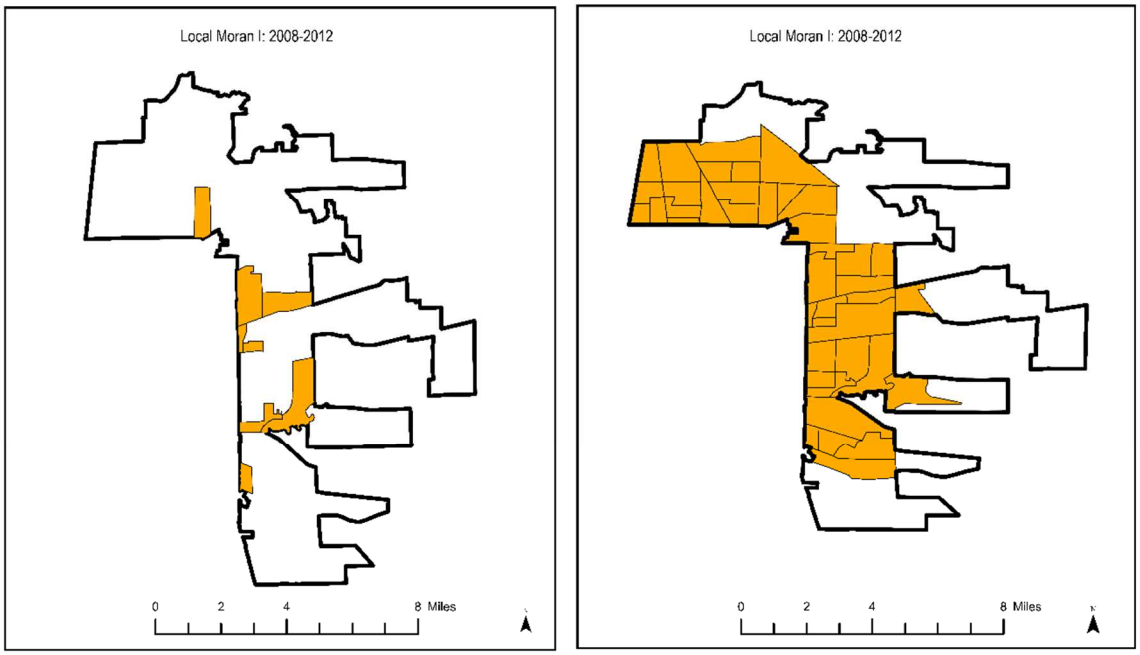
Local Moran *I* analysis of Buffalo MSA census tract poverty rates consistently reveals statistically significant concentration of higher poverty rate census tracts in the cities of the metropolitan area; specifically, the city of Buffalo, NY. The suburbanization of people in the Buffalo MSA has also begun to reveal movement of higher census tract poverty into inner ring suburban areas. Examination of the distribution of poverty census tracts using Local Moran *I* reveals a shifting distribution of high poverty census tracts in the inner ring suburban areas. In 2013-2017 and 2018-2022 study periods, clusters of high poverty census tracts began to appear in the inner ring suburban area as shown in Map 2.

Local Moran *I* analysis provides additional spatial census tract poverty distribution information that would be obscured using traditional distribution metrics such as confidence intervals highlighted in Tables 1-1B. In 2008-2012, a spatial concentration of high poverty census tracts appeared in the inner ring suburban area in the Buffalo MSA. Local Moran *I* analysis in 2018-2022 shows a decrease in the number of high-high poverty census tracts in the inner ring suburban area.

Analysis of local spatial autocorrelation should incorporate more than one spatial autocorrelation method. This paper uses an additional local spatial autocorrelation method to conduct spatial sensitivity analysis; Getis-Ord Gi*.

Getis-Ord Gi* the G-statistic (also known as Hot Spot Analysis) measures the concentration of low or high data values; a large positive *Z* value indicates a concentration of large values close to one specified neighborhood (*i*), and a large negative *Z* value implies many small values close to neighborhood *i*.

Map 2: Local Moran I and Gi* (Hot Spot Analysis): 2008-2012



Map 3: Local Moran I and Gi* (Hot Spot Analysis): 2018-2022

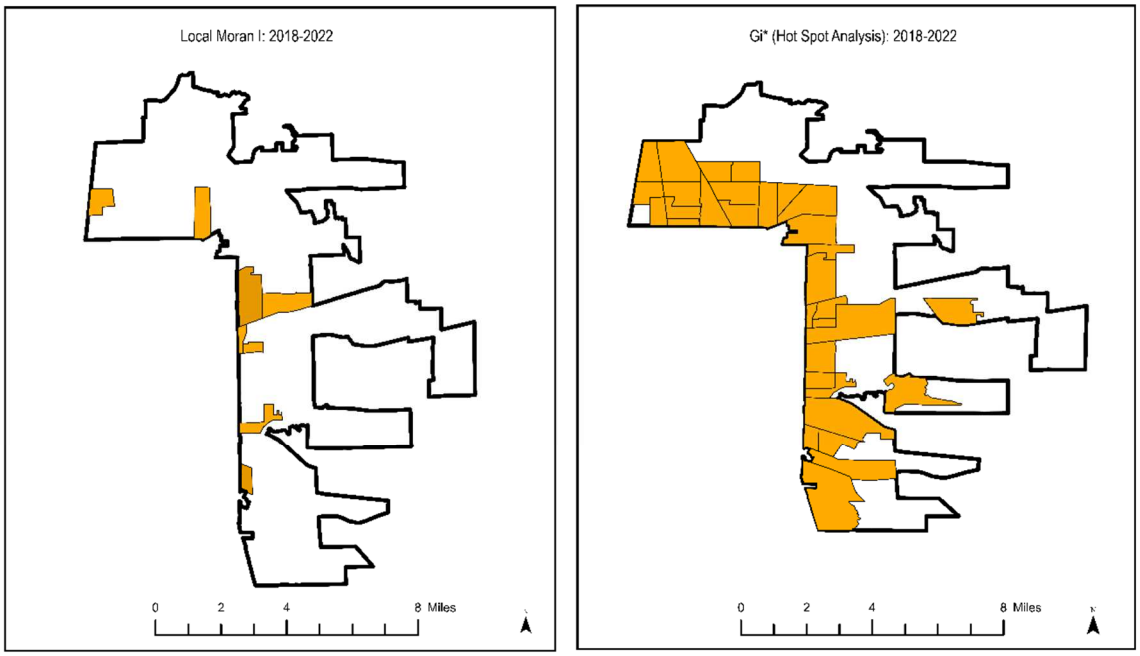


Table 2. LISA Mean Poverty Rates: Cities and Inner Ring Areas in Buffalo MSA: 2008-2022

	Local Moran *I*	Getis-Ord Gi*	Mean
Cities 2008-2012	33.77%	30.31%	25.70%
Cities 2013-2017	36.05%	27.23%	26.74%
Cities 2018-2022	32.51%	28.35%	24.04%
Inner-Ring 2008-2012	18.77%	10.36%	8.96%
Inner-Ring 2013-2017	24.28%	11.12%	9.66%
Inner-Ring 2018-2022	16.73%	10.31%	8.92%

Source: Author's computations

Getis-Ord Gi* determines hot spots (high value clusters) or cold spots (low value clusters) by including the target features values in the analysis. The difference in identification of high census tracts poverty clusters is visually striking in comparing Maps 2 and 3. Hot spot analysis identifies more census tracts in the inner ring suburban area in 2008-2012 and 2018-2022[6].

Comparison of LISA computed poverty rates in Table 2 for cities and inner ring areas show the poverty rates for cities and inner-ring areas derived from Local Moran *I* and Gi* were higher than the mean for these locales. Local Moran I and Getis -Ord Gi* analysis visually show the spread of higher poverty census tracts into non-central city areas. The mean census tract poverty rates of these high-high clusters varied by time increasing between 2008-2017 for both spatial measures and then decreasing in 2018-2022.

Across the study period 2008-2022 there were no spatially clustered high poverty census tracts in the outer ring suburban areas in the Buffalo metropolitan area identified using Gi* or Local Moran *I* analysis. The increase presence of relatively higher poverty rates and the spatial clustering of poverty in inner ring suburban areas in the Buffalo MSA should be a call to action to public policymakers, academics, and community organizations alike that intra-metropolitan initiatives and cooperations across political boundaries will increasingly need to be the foundation of initiatives and policies to reduce the incidence of intra-metropolitan poverty. Further research on the incidence of intra-metropolitan poverty will need to explore the spread of poverty within the lenses of race, gender, and age which this initial exploration did not do.

Notwithstanding the mentioned limitations of computing poverty and the need to have additional poverty measurements, using exploratory spatial data analysis within a GIS environment provides an important set of spatial tools to understand the spatial nuances of intra-metropolitan poverty analysis.

An important initiative that will need to be undertaken is to provide access to exploratory spatial data analysis training so that those organizations involved in poverty research can use this tool and spatial techniques in identifying and modeling the spread of poverty and working on development of policy initiatives to reduce the incidence of intra-metropolitan poverty.

CONCLUSION

Exploratory spatial statistics within a widely available GIS program such as ArcGIS Pro can provide policymakers and others with access to spatial information that provides spatial nuance to the examination of poverty and the distribution of poverty in metropolitan areas. It will become increasingly important for policy makers to explore policy analysis within an intra-metropolitan landscape to address the spatial distribution of poverty, the contributing factors to poverty and hopefully public and private policies to reduce the incidence of poverty in metropolitan areas such as the Buffalo metropolitan statistical area.

ENDNOTES

[1] Maps throughout this paper were created using ArcGIS® software by Esri. ArcGIS® and ArcMap™ are the intellectual property of Esri and are used herein under license. Copyright © Esri. All rights reserved. For more information about Esri® software, please visit www.esri.com.

[2] (Getis & Aldstadt, 2004) identify eleven spatial weight matrix schemes: (1) spatial contiguous neighbors; (2) inverse distances raised to some power; (3) lengths of shared borders divided by the perimeter; (4) bandwidth as the nth nearest neighbor distance; (5) ranked distances; (6) constrained weights for an observation equal to some constant; (7) all centroids within distance *d*; (8) *n* nearest neighbors; and three newer schemes : (9) bandwidth distance decay; (10) Gaussian distance decline; and (11) “tri-cube” distance decline function.

[3] There was no statistically significant difference in the ESDA analysis using either spatial weight scheme. Projected coordinated system: NAD 1983 UTM Zone 17N; Projection: Transverse Mercator

[4] ArcGIS Pro/Desktop**:** Esri Inc. (2024). *ArcGIS Pro* (Version 3.4) [Software]. Redlands, CA: ESRI, Inc.

[5] 2008-2012 Global Moran *I* index: .3126, z-score: 13.9773, p-value:0.000;2013-2017 Global Moran *I* Index: .3664 z-score: 15.9767, p-value:0.0000; and 2018-2022 Global Moran *I* Index: .3353, z-score: 16.0936, p-value: 0.000. There was no significant change between the 2008-12 and 2013-2017 LISA analysis; and thus the 2013-2017 LISA results are not shown.

[6] 2008-2012 Local Moran *I* census tracts: 93.01,102.02,101.02,103,104,109.02; Gi* census tracts: 79.02,79.03,79.04,80.01,80.02,80.03,81.01,81.02,82.02,85,86,87,88,92,93.01,93.02,94.01,95.01,99,100.03,101.01,101.02,101.03,102.01,102.02,103,104,105,106,107,108.08,109.01,109.02,110,111,113,114,115,113,117. 2018-2022 Local Moran *I* census tracts:93.01,101.02,102.02,103,104,109.02; Gi* census tracts: 79.03,79.04,80.01,80.02,80.03,81.01,81.02,82.01,82.02,85,86,87,93.01,93.02,94.01,95.01,97.01,99,102.01,102.02,103,104,105,107,108.08,110,111,113,114,115,117,120.01.

REFERENCES

Anselin, Luc. (1993). *The Moran Scatterplot as an ESDA tool to Assess Local Instability in Spatial Association.* West Virginia University, Regional Research Institute, Morgantown.

Anselin, Luc. (1995). Local Indicators of Spatial Association-LISA. *Geographical Analysis, 27*(2), 93-115.

Bellu, Lorenzo, G.and Liberati, Paolo. (2005). Impacts of Policies on Poverty: Relative Poverty Lines. MPRA Paper 44665, *University Library of Munich, Germany*, 1-13.

Bhatta, Saurav, D. (2001). Are Inequality and Poverty Harmful for Economic Growth: Evidence from the Metropolitan Areas of the United States. *Journal of Urban Affairs, 23*(3-4), 335-359.

Boots, Barry,.and Okabe, Atsuyuki.. (2007). Local Statistical Spatial Analysis: Inventory and Prospects. *International Journal of Geographical Information Science, 21*(4), 355-375.

Chetty, Raj, and Hendren, Nathaniel. (2018). The Impacts of Neighborhoods on Intergenerational Mobility I: Childhood Exposure Effects. *Quarterly Journal of Economics,* 133(3), 1107-1162

Chetty, Raj., Hendren, Nathaniell, and Katz, Lawrence. (2016). The Effects of Exposure to Better Neighborhoods on Children: New Evidence from the Moving to Opportunity Experiment. *American Economic Review,* 106(4), 855-902

Chyn, Eric, and Katz, Lawrence. (2021). Neighborhoods Matter: Assessing the Evidence for Place Effects. *Journal of Economic Perspectives,* 35(4), 197-222

Citro, Constance. F, and Michael, Robert. T. (Eds.). (1995). *Measuring Poverty: A New Approach.* Washington, D.C.: National Academy Press.

Cooke, Thmas, and Marchant, Sarah. (2006). The Changing Intrametropolitan Location of High-poverty Neighbourhoods in the US, 1990-2000. *Urban Studies,* 43(11), 1971-1989

Cushing, Brian, and Zheng, Buhong. (2000). Re-evaluating Differences in Poverty among Central City, Suburban, and Nonmetropolitan Areas of the US. *Applied Economics, 32*(5), 653-660.

Field, Andy. (2009). *Discovering Statistics Using SPSS* (Vol. 3rd). London: Sage.

Fotheringham, A. Stewart., Brunsdon, Chris, and Charlton, Martin. (2000). *Quantitative Geography: Perspectives on Spatial Data Analysis.* London: Sage.

Getis, Arthur, and Aldstadt, Jared. (2004). Constructing the Spatial Weights Matrix Using a Local Statistic. *Geographical Analysis, 35*(2), 90-104.

Getis, Arthur, and Ord, John K. (1992). The Analysis of Spatial Association by Use of Distance Statistics. *Geographical Analysis, 24*(3), 189-206.

Jolliffe, Dean. (2003). On the Relative Wellbeing of the Nonmetropolitan Poor: An Examination of Alternate of Poverty During the 1990s. *Southern Economic Journal, 70*(2), 295-311.

Johnson, Colleen, M.. (2007). A Re-count of Poverty in US Central Cities: Just Who and Where Are the Urban Poor. *Urban Studies, 44*(12), 2283-2303.

Kasarda, John, D. (1990). Structural Factors Affecting the Location and Timing of Urban Underclass Growth. *Urban Geography,* 11(3), 234-264

Kneebone, Elizabeth. (2009, October 19). *The Suburbanization of American Poverty.* Retrieved April 12, 2012, from Metropolitan Opportunity Series Number 2: http://www.brookings.edu/opinions/2009/1019_poverty_kneebone.aspx

Laliberte, Jean.W. (2021). Long-Term Contextual Effects in Education: Schools and Neighborhoods. *American Economic Journal: Economic Policy,* 13(2), 336-377

Leigh, Nancy, and Lee, Sugie.(2005). Philadelphia's Space in Between: Inner-Ring Suburb Evolution. *Opolis,* 1(1), 13-32

Li, Larry, Henry, John., and Davis, Tim. (1999). Integration of Statistics and Geography: Statistics Canada's Use of Geostatistical Data and GIS. *Statistical Journal of the UN: Economic Commission for Europe, 16*(1), 49-58.

Massey, Douglas .S,. and Denton, Nancy, .A. (1993). *American Apartheid: Segration and the Making of the Underclass.* Cambridge, MA.: Harvard University Press

Myles, John, and Picot, Garnett. (2000). Poverty Indices and Policy Analysis. *Review of Income and Wealth, 46*(2), 161-179.

Oreopoulos, Philip. (2003). The Long-Run Consequence of Living in a Poor Neighborhood. *The Quarterly Jornal of Economics, 118(4), 1533-1575.*

Orshansky, Mollie. (1965). Counting the Poor: Another Look at the Poverty Profile. *Social Security Bulletin, 28*, 3-29.

Osberg, Lars, and Xu, Kuan. (1999). Poverty Intensity: How Well Do Canadian Provinces Compare? *Canadian Public Policy, 25*(2), 179-195.

Pittman, Robert. H,,and Thrall, Grant. I. (1991). Using Geographic Systems in Economic Development: An Introduction. *Economic Development Review, 9*(4), 14-21.

Primus, Wendell. E. (2006). Reductions in Poverty Significantly Greater in the 1990s than Official Estimates Suggest. *Review of Policy Research, 23*(3), 781-797.

Sawhill, Isabel. V. (1988). Poverty in the U.S.: Why is it so Persistent. *Journal of Economic Literature*, 1073 1119.

Sen, Amartya. (1976). Poverty: An Ordinal Approach to Measurement. *Econometrica, 44*(2), 219-231.

Strait, John. (2001). The Disparate Impact of Metropolitan Economic Change: The Growth of Extreme Poverty Neighborhoods, 1970-1990. *Economic Geography,* 77(3), 272-305

U.S. Census Bureau. (2012). *Poverty Status in the Past 12 months, 2008-2012 American Community Survey 5-year estimates* [Table S1701]/ Retrieved from data.census.gov

U.S. Census Bureau. (2017). *Poverty Status in the Past 12 months, 2013-2017 American Community Survey 5-year estimates* [Table S1701]/ Retrieved from data.census.gov

U.S. Census Bureau. (2022). *Poverty Status in the Past 12 months, 2017-2022 American Community Survey 5-year estimates* [Table S1701]/ Retrieved from data.census.gov

Unwin, David, and Unwin, Antony. (1998). Exploratory Spatial Data Analysis with Local Statistics. *Journal of the Royal Statistical Society: Series D, 47*(3), 415-421L

Wilhelm, Adalbert, and Steck, Robert. (1998). Exploring Spatial Data with Interactive Graphics and Local Statistics. *The Royal statistical Society: Series D, 47*(3), 423-430.

Wilson, William. J, and Aponte, Robert. (1985). Urban Poverty. *Annual Review of Sociology, 11*, 231-258.

Wilson, William, J. (1987). *The Truly Disadvantaged: The Inner City, the Underclass, and Public Policy.* Chicago, IL.: University of Chicago Press

Wong, David. S, and Lee, Jay. (2005). *Statistical Analysis of Geographic Information.* Hoboken, NJ: John Wiley and Sons.

Diversification In Electricity Price And Volume Swaps

Matthew Brigida[1], David Brigida[2], Kathleen Maceyka[3]

ABSTRACT

This analysis determines the effect of diversification on portfolios of retail electricity price and volume swaps. These contracts are used by municipalities to provide fixed costs to customers and enable greater renewable generation. Power marketing firms often pay the floating generation cost. We find very limited benefits to diversification for power marketing firms providing these swaps within the NEPOOL region. This evidence supports a long/short swap strategy, or diversification across wide geographic regions.

JEL Codes: G12; G10
Keywords: Electricity Derivatives; Renewable Energy

INTRODUCTION

In deregulated wholesale electricity markets prices fluctuate freely given market conditions for electricity supply, demand (load) and transmission congestion. However, many purchasers of wholesale electricity, such as municipalities, prefer price certainty. Given this preference, many power marketing firms offer fixed for floating electricity swaps, often also including a given proportion of renewable generation. These swaps are analogous to other types of swaps (such as interest rate swaps) except that the swap volume is unknown. Therefore, electricity swaps allow wholesale electricity customers to pay a *fixed price* for their *uncertain electricity demand*. This is like an interest rate swap with a fluctuating notional amount. This latter aspect of electricity swaps makes them particularly difficult to value and hedge.

In this analysis we test whether power marketing firms can reduce risk for consumers by entering into fixed for floating electricity swaps across various electricity markets. That is, we estimate the effect of geographic diversification of electricity swaps. If diversification is an effective means to reduce risk, power marketing firms will be able to offer a greater number of these swaps. Thus, our analysis contributes to the literature on methods to provide stable electricity prices to end consumers, thereby facilitating economic growth.

Valuation of these swaps is an active area of research. Difficulties arise from both the unique behavior of electricity prices as well as the uncertain volume of the contract. A step toward such valuation is Nazar et al. (2018), which created a method to simultaneously forecast both power and load. Our present analysis

[1] Matthew Brigida (corresponding author), Department of Accounting and Finance, SUNY Polytechnic Institute, 100 Seymour Road, Utica, NY 13502. Email: matthew.brigida@sunypoly.edu.
[2] David Brigida, Department of Accounting and Finance, SUNY Polytechnic Institute, 100 Seymour Road, Utica, NY 13502.
[3] Kathleen Maceyka, Department of Accounting and Finance, SUNY Polytechnic Institute, 100 Seymour Road, Utica, NY 13502.

adds to the literature not by valuing individual contracts, but by determining how a portfolio of these contracts behaves. Is there a substantial diversification benefit, and over how wide of a geographic region must a power marketing firm diversify?

Anderson et al. (2007) discuss the important role forward contracts play in electricity markets. Benth and Koekebakker (2008) use a Heath-Jarrow-Morton approach to model fixed-for-floating electricity price swaps to some success. Note, in this analysis the size of the swap (megawatt hours) is fixed. Kemper et al. (2022) price derivatives based on electricity swaps in an arbitrage-free framework by introducing a market price for the delivery period of the electricity swap. Kemper and Schmeck (2023) price electricity swaps where the payoff is a geometric average. Our analysis complements the theoretical literature by using an empirical approach to focus on portfolios of contracts rather than individual contracts themselves.

Frestad et al. (2010) find evidence of substantial risk-premia in short-dated swaps on Nordic exchanges. They also find evidence that daily swap returns exhibit significant kurtosis, but little skewness. Similarly, Blanco et al. (2018) find evidence that electricity swap Value-at-Risk (VaR) estimates underestimate tail risk if a Normal distribution is used. Peña et al. (2020) determine that, at least with respect to electricity futures, decent VaR estimates can be made with a GARCH(1,1) model with Student-t errors.

Further, our present analysis contributes to risk-management and mitigation by investigating the effect of diversification in electricity swaps. In addition to lessening risk, portfolios tend to have return distributions that are closer to the Normal distribution when compared to single asset return distributions.

There is a substantial amount of research on how to incorporate intermittent renewable energy into deregulated power markets. Lai et al. (2021) investigate a portfolio hedging strategy for natural gas generators in the presence of intermittent renewable energy. Ihlemann et al. (2022) highlight the benefits of balancing power capacity markets across borders—thereby diversifying across a larger region. Zeynali et al. (2021) investigate the power price effect of coordinating wind generation with electric vehicle charging. Successful integration of renewable energy has a positive effect on regional economic growth. González and Álvarez Alonso (2021) discuss the effect of power market integration on industrial power prices.

The remainder of this paper is organized as follows. Section 2 describes the swap contract, and Section 3 covers our dataset and empirical methods. Section 4 presents results and Section 5 concludes.

SWAP DESCRIPTION

The contract is a fixed for floating power swap, and is based on contracts presently being traded by power marketing firms and municipalities. The contract has the following characteristics:

- Revenue per MWh supplied is fixed.
- Cost of supplying power is floating.
- Amount of power supplied is floating.
- Customers can *migrate* onto and away from the load that must be served.

We can write the payoff on the contract as:

$$Profit = (P - X)T_{MWh}$$

where P is the power price in \$/MWh, X is the price for which we must deliver power in \$/MWh, and T is the amount of power we must deliver in MWh. Further:

$$T_{MWh} = f(E, I)$$

where E and I denote weather and migration respectively. Migration itself is a function of the power price. If the retail power price increases, then consumers migrate to the fixed contract, and they leave the fixed contract as the price decreases.

The Effect Of Load Changes On Profit

A load increase will increase the cost of power, reducing profit per MWh. In addition there is lower profit (or a loss) on a larger amount of MWh. Thus, load increases tend to lower profit.

Alternatively, a decrease in the load will lower prices and increase profit per MWh. However this increased profit will be earned on fewer MWh, so ultimately, it is unclear whether total contract profit will increase or decrease.

Thus the contract tends to earn the most profit when load and power prices do not deviate much from those specified in the contract. The contract then has an analogue in the butterfly, and short straddle option spreads. In this analysis, we set the fixed contract price as the average price of power in the year prior and discount it by a 15% cost of capital.

DATA AND METHODS

Our analysis uses monthly Real-Time Locational Marginal Prices (RT-LMP) ranging from January 2016 to February 2022, for 74 total months. Our RT-LMP cover each region in the area overseen by the New England Independent System Operator (ISO-NE). These regions are: Maine, Connecticut, Vermont, New Hampshire, Rhode Island, South-East Massachusetts, North-East Massachusetts, and West-Central Massachusetts. Data are gathered directly from ISO New England's ISO Express data service.[1] Summary statistics of the swap contract changes calculated using each region's RT-LMP are in Table 1 below.

Table 1. Summary Statistics: Real-Time Monthly LMP Changes By Region In \$/MWh, January 2016 Through January 2022 Inclusive.

	ME	NH	VT	CT	RI	SEMA	WCMA	NEMA
Count	73.00	73.00	73.00	73.00	73.00	73.00	73.00	73.00
Mean	1.73	3.28	1.32	3.42	3.50	3.52	3.27	3.32
Std. Dev.	396.04	84.78	59.92	90.14	73.97	80.76	103.25	129.29
Minimum	-2354.74	-207.97	-195.96	-289.89	-231.78	-202.51	-330.64	-479.08
25%	-33.82	-38.49	-29.90	-27.09	-42.67	-40.24	-35.88	-45.76
50%	1.17	0.11	-1.42	0.64	0.86	-0.88	0.44	5.79
75%	21.52	24.86	30.71	38.02	43.52	36.81	44.46	39.89
Maximum	2322.53	207.61	187.38	295.33	254.35	298.51	337.05	606.29

Notes: ME = Maine, NH = New Hampshire, VT = Vermont, CT = Connecticut, RI = Rhode Island, SEMA = South-East Massachusetts, WCMA = West-Central Massachusetts, NEMA = North-East Massachusetts.

Table 2. Correlation Matrix: Real-Time Monthly LMP Changes.

	ME	NH	VT	CT	RI	SEMA	WCMA	NEMA
ME	1.00	0.09	0.05	0.02	0.05	0.06	0.05	0.03
NH	0.09	1.00	0.72	0.62	0.57	0.50	0.55	0.49
VT	0.05	0.72	1.00	0.41	0.45	0.35	0.37	0.41
CT	0.02	0.62	0.41	1.00	0.78	0.61	0.92	0.85
RI	0.05	0.57	0.45	0.78	1.00	0.70	0.77	0.69
SEMA	0.06	0.50	0.35	0.61	0.70	1.00	0.65	0.58
WCMA	0.05	0.55	0.37	0.92	0.77	0.65	1.00	0.90
NEMA	0.03	0.49	0.41	0.85	0.69	0.58	0.90	1.00

Notes: Region abbreviations are defined in the notes to Table 1.

The correlation coefficients between ISO-NE locations range from 0.02 between Connecticut and Maine, to 0.92 between Connecticut and West-Central Massachusetts. Generally, Maine has the smallest correlation with the rest of New England. This makes Maine attractive for diversification, however Maine also has the highest volatility.

Results from Jarque-Bera tests for Normality are in Table 3 below. Over each region we reject the null, which is evidence that the price difference series do not have skewness and kurtosis consistent with the Normal distribution. Notably, however, electricity returns (over periods with positive prices) also generally fail Normality tests, as do individual stock returns. Additionally, mean-variance optimization is appropriate with non-Normal returns assuming market participants have quadratic utility—that is, their utility functions do not include 3rd or higher moments.

Table 3. Jarque-Bera Test For Normality Results.

	ME	NH	VT	CT	RI	SEMA	WCMA	NEMA
JB statistic	2972.29	4.67	12.84	28.05	9.19	17.00	36.42	203.31
p-value	0.00	0.10	0.00	0.00	0.01	0.00	0.00	0.00

Notes: Region abbreviations are defined in the notes to Table 1.

RESULTS

Single Zone Results

Below we provide results for each zone, and long-only and long-short portfolios across all zones. We use the Sharpe Ratio to evaluate portfolio performance because it is the most widely used risk-adjusted measure.

Table 4. Contract Percentage Profit, Individual Zone Data.

Zone	Mean Return	Standard Deviation	Sharpe Ratio
Maine	1.11	1.09	1.02
New Hampshire	0.72	0.15	4.64
Vermont	0.76	0.25	3.04
Connecticut	0.70	0.23	3.09
Rhode Island	0.69	0.17	3.95
SE Massachusetts	0.76	0.26	2.99
WC Massachusetts	0.71	0.17	4.24
NE Massachusetts	0.75	0.38	2.00
Average	0.78	0.34	3.12

Notes: Mean returns and standard deviations are stated as decimal fractions. The Sharpe Ratio is the mean return divided by the standard deviation of returns. The final row is the equally weighted average across the eight zones.

The Sharpe Ratio for Maine is the lowest at 1.02. The Sharpe Ratio for New Hampshire is the highest at 4.64. These Sharpe Ratios are relatively high and reflect our construction of the contract—Sharpe Ratios would likely be lower if we did not reset the contract costs after each year.

Portfolio Results

We construct both long-only and long-short portfolios. Typically power marketing firms take the long side of these contracts, and municipalities and other sources of load take the short side. That said, through trading with other power marketing firms, it is possible for a firm to take a short position. It may be realistic, however, to limit the amount which can be sold.

We calculate mean-variance optimal portfolios by maximizing the Sharpe ratio (using Martin, 2021). This method assumes quadratic utility or asset returns are Normally distributed (or a member of the elliptical family of distributions). That is, we assume investors are indifferent to the third, and higher moments of the return distribution. This assumption is common when dealing with stock portfolios, despite evidence that stock returns may exhibit significant skewness and kurtosis. The assumption is likewise troublesome when applied to electricity returns and the returns on our swap contracts.

All Zone Portfolio

Below are results for long-only and long-short portfolios. A long-only portfolio is more consistent with the role of a power-marketing firm, however these firms are able to enter the short side of the swap contract (buying rather than supplying power).

No Shorting

Our long-only portfolio has a Sharpe ratio of 4.58, which is slightly less than the maximum Sharpe ratio across individual regions (New Hampshire has a Sharpe ratio of 4.64). It is, however, greater than the average Sharpe Ratio over all regions (3.12). Somewhat counterintuitively, mean-variance optimization allocated most of the portfolio to West-Central Massachusetts and Rhode Island, with a 0% weight to the

region with the highest Sharpe Ratio (New Hampshire). Again these are high Sharpe ratios and should only be compared to each other given the assumptions made in the contract construction. Portfolio performance is reported in Table 5, and the region weights in Table 6.

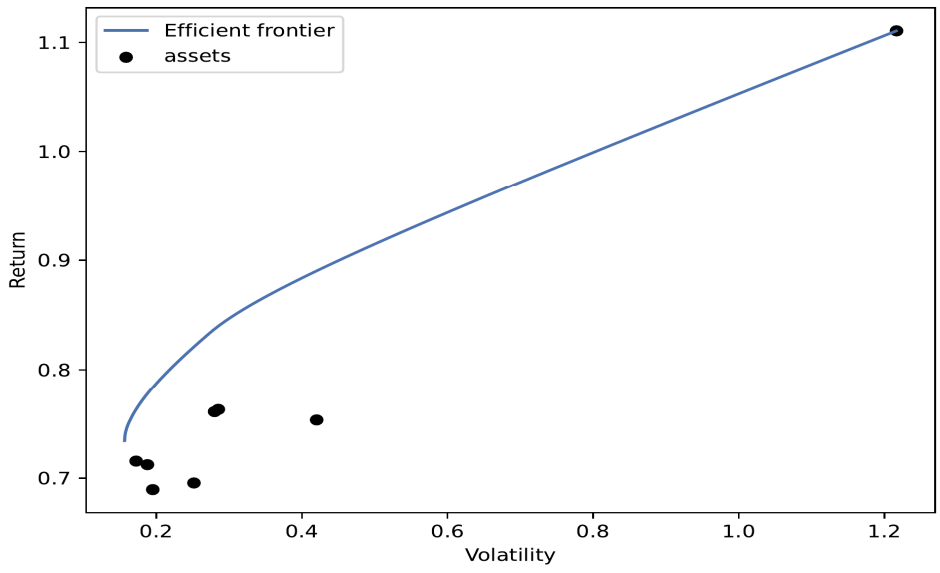
Table 5. All Zone Portfolio Performance, Long-Only Portfolio (Weights Unconstrained).

Statistic	Value
Expected Return	73.94%
Standard Deviation	15.70%
Sharpe Ratio	4.58

Table 6. Region Weights In The Long-Only Portfolio (Weights Unconstrained).

Region	Weight
Maine	8.09%
New Hampshire	0.00%
Vermont	0.00%
Connecticut	0.00%
Rhode Island	22.21%
SE Massachusetts	0.00%
West-Central Massachusetts	69.70%
NE Massachusetts	0.00%

Figure 1. Efficient Frontier For The Long-Only Portfolio.



Notes: The solid line is the efficient frontier constructed from the eight ISO-NE zone swap contracts. Points denote the individual zone contracts.

Long-Only Weights Constrained To 20%

In the above long-only portfolio 5 regions have 0% weights, and approximately 70% of the portfolio is in West-Central Massachusetts. Since many power marketing firms have internal rules regarding maximum exposure to any given region or contract, it makes sense to additionally restrict the weight of each region. In Tables 7 and 8 we provide a long-only optimal portfolio with the maximum weight in any region constrained to 20%.

Constraining the weights to a maximum of 20% lowers the Sharpe Ratio to 3.95 from 4.58 in the long-only portfolio. The reduction in the Sharpe Ratio is due to an increase in the portfolio standard deviation from 15.70% to 18.47%. The expected return increased with the additional constraints. With the additional 20% constraint, only two regions (Vermont and NE Massachusetts) receive a 0% weight.

Table 7. All Zone Portfolio Performance, Long-Only Portfolio (Weights Constrained To 20%).

Statistic	Value
Expected Return	74.99%
Standard Deviation	18.47%
Sharpe Ratio	3.95

Table 8. Region Weights In The Long-Only Portfolio (Weights Constrained To 20%).

Region	Weight
Maine	8.35%
New Hampshire	20.00%
Vermont	0.00%
Connecticut	11.65%
Rhode Island	20.00%
SE Massachusetts	20.00%
West-Central Massachusetts	20.00%
NE Massachusetts	0.00%

Shorting Allowed

In the two portfolios below we allow the power marketing firm to take the short side of the contract. Given the generally high correlations between the contract returns across regions, allowing shorting will allow the portfolio standard deviation to be greatly reduced without lowering the mean return. That said, traditionally power marketing firms are on the long side of these types of contracts, because they are generally an arm of a power producer. The short side is typically a municipality or other load source. Therefore, the ability to take the short side of these contracts may be limited.

Weights Constrained To 100%

Allowing up to 100% short positions results in a very low portfolio standard deviation and high Sharpe Ratio at 1.75% and 34.39 respectively. The optimization took advantage of the high positive correlations and allocated negative weights of approximately -82%, -92%, and -100% to New Hampshire, South-East

Massachusetts, and North-East Massachusetts respectively. This is evidence that geographically constrained power marketing firms should consider shorting these swap contracts to the extent that it is allowed by the market and corporate policy.

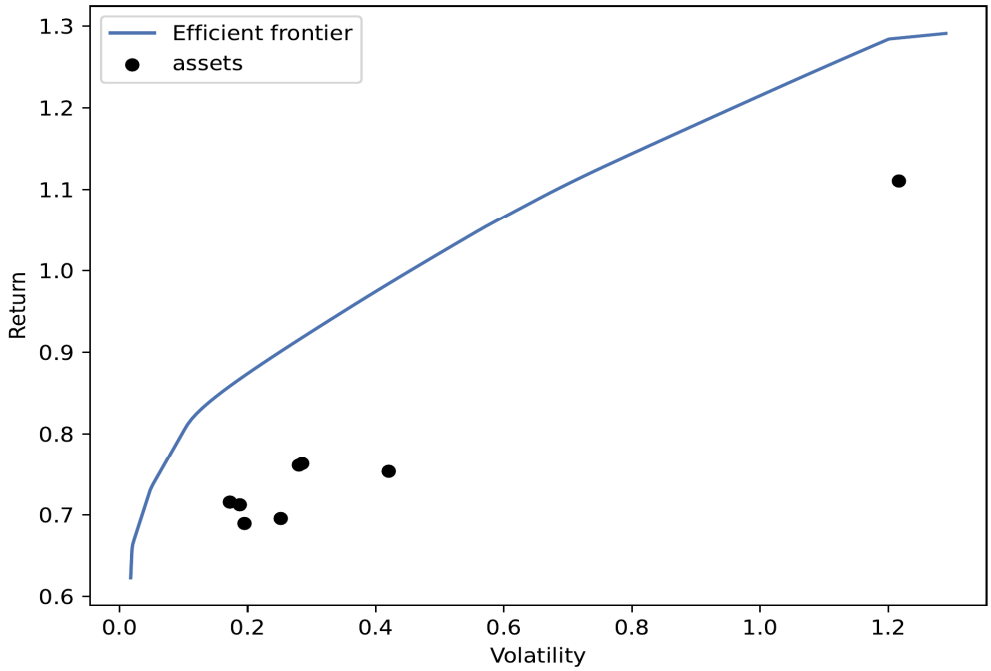
Table 9. All Zone Portfolio Performance, Full-Shorting-Allowed Portfolio (Weights Constrained To 100%).

Statistic	Value
Expected Return	62.30%
Standard Deviation	1.75%
Sharpe Ratio	34.39

Table 10. Region Weights In The Shorting-Allowed Portfolio (Weights Constrained To 100%).

Region	Weight
Maine	1.39%
New Hampshire	-82.03%
Vermont	72.57%
Connecticut	100.00%
Rhode Island	100.00%
SE Massachusetts	-91.93%
West-Central Massachusetts	100.00%
NE Massachusetts	-100.00%

Figure 2. Efficient Frontier For The Full-Shorting-Allowed Portfolio.



Notes: The solid line is the efficient frontier constructed from the eight ISO-NE zone swap contracts when weights may range from -100% to 100%. Points denote the individual zone contracts.

Weights Constrained To 20%

It is more reasonable for a power marketing firm to constrain the firm's exposure to a given contract and market. The portfolio below constrains the weight in each region to between -20% and 20%. When doing so the portfolio's Sharpe Ratio of 4.43 is slightly *lower* than the Sharpe Ratio of the long-only (but otherwise unconstrained) portfolio (4.58). The only region shorted in the portfolio is North-East Massachusetts, which had the lowest individual region Sharpe Ratio after Maine.

Table 11. All Zone Portfolio Performance, Shorting-Allowed Portfolio (Weights Constrained To 20%).

Statistic	Value
Expected Return	73.96%
Standard Deviation	16.22%
Sharpe Ratio	4.43

Table 12. Region Weights In The Shorting-Allowed Portfolio (Weights Constrained To 20%).

Region	Weight
Maine	6.52%
New Hampshire	20.00%
Vermont	13.48%
Connecticut	20.00%
Rhode Island	20.00%
SE Massachusetts	20.00%
West-Central Massachusetts	20.00%
NE Massachusetts	-20.00%

CONCLUSION

In this analysis we have calculated returns on price and volume swap contracts which are often provided by power marketing firms to municipalities. While these contracts may have a high Sharpe Ratio individually, we have found evidence for the limited diversification potential on the New England ISO for these swaps. For a power marketing firm to limit risk within the ISO, they would have to construct long-short power swap portfolios. This essentially means risk can only be allocated among similar firms, and therefore does not reduce the risk held by the overall market participants.

This highlights the need for power marketing firms to be able to diversify across geographical markets. This is often limited by differing power market structure and corporate policy. Therefore, we find support for measures which allow power marketing firms to operate with greater geographic diversity. However, a remaining question for future analyses is the extent to which greater geographic diversification can reduce risk. Specifically, measuring the effect of adding contracts on the ERCOT (Texas) or CAISO (California) operated electricity grids is a sensible area of future research.

ACKNOWLEDGMENTS

CRediT authorship contribution statement. David Brigida: Methodology, Software, Data curation, Investigation, Formal analysis, Writing – original draft. Kathleen Maceyka: Conceptualization, Software,

Writing – original draft, Writing – review & editing. Matthew Brigida: Conceptualization, Methodology, Writing – original draft, Writing – review & editing.

ENDNOTES

1. ISO New England ISO Express data service, https://www.iso-ne.com/markets-operations/iso-express.

REFERENCES

Anderson, E. J., Hu, X., and Winchester, D. (2007). "Forward Contracts in Electricity Markets: The Australian Experience." *Energy Policy*, 35(5), 3089–3103.

Benth, F. E., and Koekebakker, S. (2008). "Stochastic Modeling of Financial Electricity Contracts." *Energy Economics*, 30(3), 1116–1157.

Blanco, I., Peña, J. I., and Rodriguez, R. (2018). "Modelling Electricity Swaps with Stochastic Forward Premium Models." *The Energy Journal*, 39(2).

Frestad, D., Benth, F. E., and Koekebakker, S. (2010). "Modeling Term Structure Dynamics in the Nordic Electricity Swap Market." *The Energy Journal*, 31(2).

González, J. S., and Álvarez Alonso, C. (2021). "Industrial Electricity Prices in Spain: A Discussion in the Context of the European Internal Energy Market." *Energy Policy*, 148, 111930.

Ihlemann, M., van Stiphout, A., Poncelet, K., and Delarue, E. (2022). "Benefits of Regional Coordination of Balancing Capacity Markets in Future European Electricity Markets." *Applied Energy*, 314, 118874.

Kemper, A., and Schmeck, M. D. (2023). "Pricing of Electricity Swaps with Geometric Averaging." *Center for Mathematical Economics Working Papers*, Bielefeld University.

Kemper, A., Schmeck, M. D., and Balci, A. Kh. (2022). "The Market Price of Risk for Delivery Periods: Pricing Swaps and Options in Electricity Markets." *Energy Economics*, 113, 106221.

Lai, S., Qiu, J., Tao, Y., and Zhao, J. (2021). "Risk Hedging for Gas Power Generation Considering Power-to-Gas Energy Storage in Three Different Electricity Markets." *Applied Energy*, 291, 116822.

Martin, R. A. (2021). "PyPortfolioOpt: Portfolio Optimization in Python." *Journal of Open Source Software*, 6(61), 3066.

Nazar, M. S., Fard, A. E., Heidari, A., Shafie-khah, M., and Catalão, J. P. S. (2018). "Hybrid Model Using Three-Stage Algorithm for Simultaneous Load and Price Forecasting." *Electric Power Systems Research*, 165, 214–228.

Peña, J. I., Rodríguez, R., and Mayoral, S. (2020). "Tail Risk of Electricity Futures." *Energy Economics*, 91, 104886.

Zeynali, S., Nasiri, N., Marzband, M., and Ravadanegh, S. N. (2021). "A Hybrid Robust-Stochastic Framework for Strategic Scheduling of Integrated Wind Farm and Plug-in Hybrid Electric Vehicle Fleets." *Applied Energy*, 300, 117432.

Comparative Analysis of Inflation Forecasting using Machine Learning: Evidence from Selected G20 Countries

Rolando A. Santos[1], Brian Sloboda[2]

ABSTRACT

With the recent trends in global inflation, much of the current research focuses on forecasting its future trajectory. This paper analyzes contemporary inflation trends using data from selected G20 countries and applies machine learning methodologies, specifically XGBoost and Facebook (also known as Prophet), to forecast future inflation. The study compares the performance and efficiency of both models, highlighting the strengths and limitations of each. While Prophet provides a valuable decomposition of historical inflation data, revealing seasonal spikes during specific months, its forecasting accuracy is less robust than that of XGBoost. XGBoost consistently outperformed Prophet, showing a significantly lower Root Mean Square Error (RMSE), particularly after hyperparameter tuning. Both models suggest a slow upward trend in inflation for the short-run out-of-sample forecast; however, the more reliable projections are drawn from XGBoost. The conclusions are aligned with the consensus on short-term inflation expectations for the selected countries.

INTRODUCTION

Inflation forecasts are crucial for private sector and central bank decision-making. Policymakers must decide how to address inflationary or deflationary pressures in advance, as monetary policy affects the economy through transmission processes, often with a significant lag (Szafranek & Polski, 2017). Numerous studies have been conducted with challenges that have accurately predicted future movements in macroeconomic variables, particularly inflation. Several important studies have examined the relative accuracy of inflation forecasts. Stock and Watson (2007) found that simple univariate models produce strong comparative inflation forecasts for the United States, and models of unobserved components and stochastic volatility produce reasonable estimates of inflation forecasts. Similarly, Faust and Wright (2013) used opinions to estimate inflation, and opinions, such as those from the Federal Reserve or surveys of inflation expectations, provide accurate estimates of inflation. More importantly, it is not surprising that these results mostly pertain to a developed economy's stable economy, given that the conditional mean of inflation was extremely persistent (Fuhrer, 2010; Wolters & Tillmann, 2015). On the other hand, if an economy is experiencing volatility, these models may need to account for long-term deviations from steady-state values. That is, a prediction near the historical mean would not be very accurate. As a result, low- and middle-income nations may need to recalibrate their models to account for specific economic conditions such as declines or higher inflation.

Inflation has been a focus of recent news around the world due to the global pandemic and other factors that have contributed to it. In the US, recent news about inflation has been considered as good news as it

[1] Department of Economics, John Carroll University, 1 John Carroll Blvd, University Heights, OH 44118
[2] Department of Economics, University of Maryland, Global Campus, 3501 University Blvd. East, Adelphi, MD 20783

recently decreased from a high of 9% to 2.4% (April 2025). In other countries, persistent inflation has been a more significant problem, leading to regional slowdowns, such as in the European Union and certain Asian countries. This paper aims to examine the future of global inflation in various selected G20 countries. We looked at some trends and compared the forecasts of each country. By utilizing recent machine learning technologies, two tools will be employed: XGBoost and Prophet. We examined the efficacy of each method and outlined the advantages and disadvantages of using each approach.

THE EXISTING LITERATURE

Since Prophet and XGBoost are recent machine learning programs, there is a limited number of studies that have been done for inflation forecasting. Menculini et al. (2021) demonstrated that various machine learning techniques, including ARIMA, LSTM, and Prophet, were utilized to compare sales price forecasts generated by an Italian food wholesaler, as a step towards automating pricing tasks typically performed by a human workforce. The results indicate that ARIMA models and LSTM neural networks perform similarly for the forecasting task under consideration. In contrast, the combination of LSTM (long-short term memory) and CNNs (Convolutional Neural Networks) performs the best accuracy. However, Prophet performed the least accurately but the most efficiently and easily (Menculini, 2021).

For neural networks analysis, Purnawansyah et al (2019) presented a prediction of the inflation rate in Samarinda, East Kalimantan, Indonesia, using an intelligent algorithm known as Backpropagation Neural Networks (BPNN). The results showed that the BPNN algorithm can be used as an alternative method in predicting inflation rates to support sustainable economic growth and improve the welfare of the people of Samarinda, East Kalimantan. In the study, inflation stability is a prerequisite for sustainable economic growth, so it is expected to provide benefits to improve people's welfare. In a similar paper on neural networks, Choudhary & Haider (2012) used a combination of neural networks to forecast the monthly inflation rates of 28 OECD member nations. In 45 percent of the nations, neural network models had superior average performance, whereas in 23 percent, AR1 performed better. Ultimately, they propose combining several network types into an ensemble to enhance the accuracy of these forecasting models.

Zahara & Ilmiddaviq (2020) utilized Long Short-Term Memory (LSTM), a specialized type of recurrent neural network (RNN), to predict inflation in Indonesia based on the exogenous daily online prices of thirty-four food staples. They employed Recurrent Neural Networks (RNNs) to forecast the Consumer Price Index (CPI). Their empirical results indicated that the LSTM model, particularly when optimized with Nesterov-accelerated Adaptive Moment Estimation (Nadam) resulted in a Root Mean Square Error (RMSE) of 4.088, which outperforms machine learning methods used in their study. Using the LSTM, Almosova & Andresen (2023) forecasted inflation. Random walk, seasonal autoregressive, Markov switching, and fully connected neural network models were some of the baseline models compared. The LSTM results, compared to the baseline models used in their analysis, showed that the root-mean-squared forecast was more accurate than the other baselines.

Garcia et al (2017) utilized machine learning (ML) methods in 12 different periods for inflation in Brazil. The first forecast period is five days before the release of inflation statistics, and the twelfth forecast period is 11 months and five days before the release of inflation statistics. In their work, the authors utilize various macroeconomic variables (a total of 59 series), which reflect the state of the financial market, labor market, balance of payments, and public debt of the country. Models are trained on a nine-year rolling window; the RMSE and the MAE serve as criteria for the quality of the forecast. The authors used models: random-walk, AR(p), LASSO, Random Forest as part of their analysis to compare different models. Their results indicate that the Random Forest and the dynamic factor model perform better than the AR and RW (random walk) model for overall forecasting horizons.

Another empirical approach in machine learning is the contraction or the shrinkage approach, which the authors used to estimate inflation rates for the Turkish economy (Özgür & Akkoç, 2022). Their empirical results show that contraction approaches are superior in terms of variable selection, and the LASSO and elastic net algorithms outperform econometric methods for Turkish inflation. These algorithms evaluate energy production factors, a measure of the construction industry, the real effective exchange rate, and indicators of the money market to identify which variables are most relevant for inflation predictions. Baybuza (2018) used boosting and random forest models to estimate inflation in Russia, and his empirical results outperform traditional models, such as random walk and autoregressive methods. His study utilized support vector machines (SVMs), K-nearest neighbors (KNNs), random forests, neural networks (ANNs), gradient boosting, and decision trees.

Medeiros et al. (2021) explored the application of machine learning techniques to forecast U.S. inflation using hundreds of macroeconomic variables from the FRED-MD database. They assessed various machine learning (ML) models, including Random Forests, LASSO, Ridge Regression, and ensemble methods such as bagging and boosting, comparing their performance against traditional models like autoregressive (AR) models and factor models. They found that machine models, particularly Random Forests, outperform traditional models in forecasting inflation across different horizons. More importantly, they found that random forests could capture the nonlinear relationships between various macroeconomic indicators and inflation that traditional linear models miss. Like Medeiros et al. (2021), Rodriguez-Vargas (2020) compared the performance of machine learning models against traditional univariate forecasting methods employed by the Central Bank of Costa Rica. The Long Short-Term Memory (LSTM) neural network outperformed other models, providing unbiased and precise forecasts, especially for horizons longer than one month. The univariate K-Nearest Neighbors (KNN) model showed strong performance, followed by random forests and extreme gradient boosting, though to a lesser extent.

The use of neural networks to forecast macroeconomic variables started in the 1990s, with the first developments on the topic dating from the 1940s (McCulloch & Pitts, 1943; Hebb, 1949). The reason for this rising popularity is that ANN pays particular attention to nonlinearities and learning processes, both of which can help improve predictions for complex variables. The following are some of the early works that utilize neural networks in macroeconomics. Tkacz and Hu (1999) forecasted Canada's GDP, while Stock

and Watson (1998) compared the forecasting performance of univariate linear and non-linear models for several macroeconomic series, finding that artificial neural networks performed poorly relative to linear models.

Moshiri and Cameron (2000) forecasted inflation using neural networks, with 25 years of monthly data available to them, which yields only 300 observations for the United States. Nakamura (2005) demonstrated that applying early stopping close to local minima improves forecasting ability, a finding that had already been suggested by Gonzalez (2000). Smalter Hall and Cook (2017) forecasted employment indicators using several deep neural network architectures, including LSTM, and showed that these improve accuracy for simpler configurations. However, they did not compare these forecasts to machine learning (ML) methods but instead compared them only to other neural network models.

In brief, the studies discussed in this literature review regarding inflation forecasting reveal the accuracy of various econometric models in predicting inflation in developed and developing countries. These studies found that machine-learning models are more accurate. Despite the abundance of relevant exogenous elements, the limitation of twelve-monthly observations per year restricts the selection of training instances. Statistical metrics exhibit temporal changes, such as the average inflation rate, conditional volatility, and persistence levels.

METHODOLOGY AND DATA SOURCES

Data Sources

The data included the following variables: Real GDP, industrial production, unemployment rate, retail sales, and federal funds rate, all measured monthly from January 1995 through May 2024. We chose time period of January 1995 to May 2024 for the following reasons:

1. Comprehensive Economic Cycles: This period encompasses various significant economic cycles, including periods of growth, recession, and recovery. Analyzing inflation during these diverse phases allows for a more robust understanding of inflation dynamics and the factors influencing them.
2. Major Global Events: The selected time frame includes critical global events, such as the Asian Financial Crisis (1997), the global financial crisis of 2008, and the COVID-19 pandemic. These events had profound effects on inflation and economic conditions worldwide, making it essential to include them in the analysis to comprehend their long-term implications.
3. Data Availability and Quality: The selection also reflects the availability of reliable and consistent data for the various economic indicators utilized in the study. This ensures the accuracy and validity of the forecasting models applied.
4. Recent Trends: By extending to May 2024, the study captures the most recent trends in inflation, particularly as countries continue to recover from the impacts of the pandemic and address inflational pressures in a post-pandemic context. This helps in generating relevant insights for policymakers and stakeholders in the current economic climate.

The data were obtained from Trading Economics (https://tradingeconomics.com/). We used XGBoost and Prophet for forecasting purposes and evaluated their performance as forecasting tools for countries like the United States, Germany, China, and Australia. We selected the G20 to include a diverse mix of both developed and emerging economies. This diversity allows for a comprehensive analysis of inflation dynamics across varying economic structures, monetary policies, and stages of development. By focusing on G20 countries, the our leverages the rich contrast in inflationary behaviors and responses to economic shocks.

Boosting Methods

The data sources used in this model are the following: Inflation rate, GDP rates, interest rates, and unemployment rates for six selected G20 countries (Argentina, China, Japan, UK, USA, and Germany). This paper used two machine learning methods: XGBoost and FB Prophet. XGBoost is considered one of the most efficient and best forecasting methods. In contrast, Prophet is efficient, especially in capturing the trend and seasonal effects of the sample.

The first method used is Extreme Gradient Boosting, also known as XGBoost. XGBoost is a powerful and robust machine learning algorithm that belongs to the family of gradient boosting algorithms. It is known for its efficiency, scalability, and high predictive performance. In recent years, XGBoost has gained popularity and achieved state-of-the-art results in various machine-learning competitions and real-world applications. Here are some key features and concepts associated with XGBoost (Chen et al., 2015; Chen & Guestrin, 2016).

- Gradient Boosting: XGBoost is based on the gradient boosting framework, which combines an ensemble of weak prediction models (typically decision trees) to create a strong predictive model. Gradient boosting iteratively adds new models to correct the errors made by previous models. The models are trained sequentially, with each new model focusing on the residuals (the differences between the predicted and true values) of the previous models.
- Regularization: XGBoost includes regularization techniques to prevent overfitting. Regularization penalizes the loss function during training to discourage the model from becoming too complex. This helps generalize to unseen data and improve the model's performance.
- Gradient-Based Optimization: XGBoost employs a gradient-based optimization algorithm to efficiently minimize the loss function and determine the optimal model parameters. It computes gradients and uses them to update the model, thereby minimizing the loss. The optimization algorithm also incorporates second-order derivatives to accelerate convergence and improve accuracy.
- Ensemble of Trees: XGBoost builds an ensemble of decision trees, where each tree captures different patterns and contributes to the final prediction. Decision trees are typically shallow, with few levels (low depth), to reduce overfitting and improve interpretability. XGBoost also supports various types of base learners, including linear models and decision trees.

• Parallelization and Scalability: XGBoost is designed to be highly scalable and can efficiently handle large datasets. It supports parallelization, both during the training of individual trees and during the feature selection process, enabling faster training times, especially on systems with multiple processors or in distributed computing environments.

Methodology of Prophet

Prophet is a forecasting algorithm developed by Facebook's data science team in 2017. The algorithm is designed to be scalable, fast, and accurate, making it suitable for a wide range of applications, from predicting sales in e-commerce to forecasting weather patterns. It is built on Stan, a probabilistic programming language that enables Bayesian inference. The model is not explicitly written as a Generalized Additive Model (GAM), but Prophet is a type of GAM. The GAM is a prior placed on parameters and updated using observed data, allowing for the quantification of uncertainty and flexible trend modeling. The general model for GAM is

$y(t)=g(t)+s(t)+h(t)+\epsilon_t$

where:

- y(t) is the observed time series data,
- g(t) represents the trend component,
- s(t) accounts for seasonality,
- h(t) captures holiday effects,
- ϵ_t is the error term.

The trend component captures the overall direction of the time series, whether it is increasing or decreasing over time. This component is modeled using a piecewise linear regression model, which allows for flexibility in fitting the trend to the data. The seasonality component captures the periodic patterns in the data, such as weekly or monthly trends. This component is modelled using the Fourier Series, allowing for flexible modelling of different seasonal patterns.

The noise component captures the random fluctuations in the data that the trend or seasonality components cannot explain. The Prophet uses a trend equation composed of multiple connected straight-line segments, allowing the slope to change at specific points in time, known as changepoints. The trend equation is

$$g(t) = \left(\left(k + \sum_{j=1}^{S} a_j(t)\delta_j\right)t + \left(m + \sum_{j=1}^{S} a_j(t)(-s\delta_j)\right)\right)$$

This trend equation works as follows. Before the first changepoint, the trend is linear with slope k. Then at each changepoint, s_j: the slope changes by δ_j, so the total slope becomes $k\sum\delta_j k$. The second sum (offset correction) ensures the continuity of the trend line. That is, the function stays smooth despite the changes in the slope.

Prophet uses a Bayesian framework to model the time series data. These equations allow Prophet to model complex time series patterns while incorporating Bayesian priors for parameter estimation. By doing so, the algorithm can generate probabilistic forecasts that provide a measure of uncertainty around the point forecast (Zunic et al, 2020).

As mentioned, Prophet uses Bayesian inference to estimate the posterior distribution of the model parameters using a Markov Chain Monte Carlo (MCMC) algorithm.[i] The Markov Chain Monte Carlo samples the posterior distribution of the model parameters. The algorithm estimates the posterior distribution of the model parameters using a likelihood function:

$$p(y|\theta)=\Pi(t=1 \text{ to } T)N(y(t)|g(t),\sigma$$

where y(t) is the observed value of the time series at time t, g(t) is the predicted value of the time series at time t, σ is the noise parameter, and Θ is the set of model parameters (including the trend and seasonality components). The algorithm uses the likelihood function to compute the posterior distribution of the model parameters using Bayes' Theorem:

$$p(\theta|y) = p(y|\theta) * \frac{p(\theta)}{p(y)}$$

where p(Θ|y) represents the posterior distribution of the model parameters, p(y|Θ) denotes the likelihood function, p(Θ) denotes the prior distribution of the model parameters, and p(y) denotes the marginal likelihood (Khare, 2023).

Prophet, an open-source software developed by Facebook's Core Data Science team, is a procedure designed for forecasting time series data based on an additive model that fits nonlinear trends with yearly, weekly, and daily seasonality, as well as holiday effects. It works best with time series that have strong seasonal effects and several seasons of historical data. The Prophet model is robust to missing data and shifts in the trend and typically handles outliers well. In this paper, Facebook's Prophet tool is used to model the dynamics of inflation, with and without the use of additional regressors, to generate monthly inflation forecasts. It is worth mentioning that an empirical method is employed to incorporate domain knowledge into the proposed framework, utilizing a method for adjusting model parameters (Zunic et al., 2020).

Rationale for Using XGBoost and Facebook Prophet in this Paper

Facebook Prophet was the main driver of this analysis despite the fact there are other ML methodologies available. More importantly, Facebook Prophet was intentionally chosen for its usability and effectiveness in handling time series forecasting. Facebook Prophet's design allows it to utilize the available data without requiring imputation or complete datasets, which is useful when analyzing inflation trends that may have irregularities in reporting. This robustness enables us to generate forecasts even when faced with incomplete datasets, ensuring that the predictions remain reliable.

Another significant feature of Prophet is its ability to generate uncertainty intervals for forecasts, which provide a probabilistic interpretation of its predictions. These confidence intervals are crucial for inflation forecasting, where policymakers and businesses must make decisions under uncertainty.

Additionally, the Bayesian framework used by Facebook Prophet enhances this capability by allowing the model to incorporate prior distributions and the variability in the data. The latter adds a layer of sophistication in capturing uncertainties that may derive from external shocks or changing economic conditions.

Evaluation of the Forecasts

We used three performance metrics for measuring forecasting accuracy for each sample.

1. Mean Squared Error (MSE)
2. Root Mean Square Error (RMSE_
3. Mean Absolute Percentage Error. (MAPE)

The mean squared error (MSE) is a measure of the average squared forecast error that avoids the problem of positive and negative forecast errors offsetting each other:

MSE = $\sum_{t=k+1}^{n} e_t^2$/ n-k

The root mean square error (RMSE) is another measure that addresses the issue of positive and negative errors canceling each other out. It is obtained by computing the square root of the average of the squared forecast errors.

RMSE = $\sqrt[2]{\frac{\sum_{t=k+1}^{n} e_t^2}{n-k}}$

The results of each country are compared across forecasting methods. The lowest forecast error will be used to calculate the out-of-sample forecast for each country. Since the regressors are unknown for future forecasts, our forecasts are based on a univariate case. The parameters and coefficients are irrelevant in FBPropet as what is more important is the overall evaluation metrics such as the mean square error(MSE) and the root mean square error (RMSE). Furthermore, the actual out of data forecast will also be presented in a graphical form to compare it with the actual results.

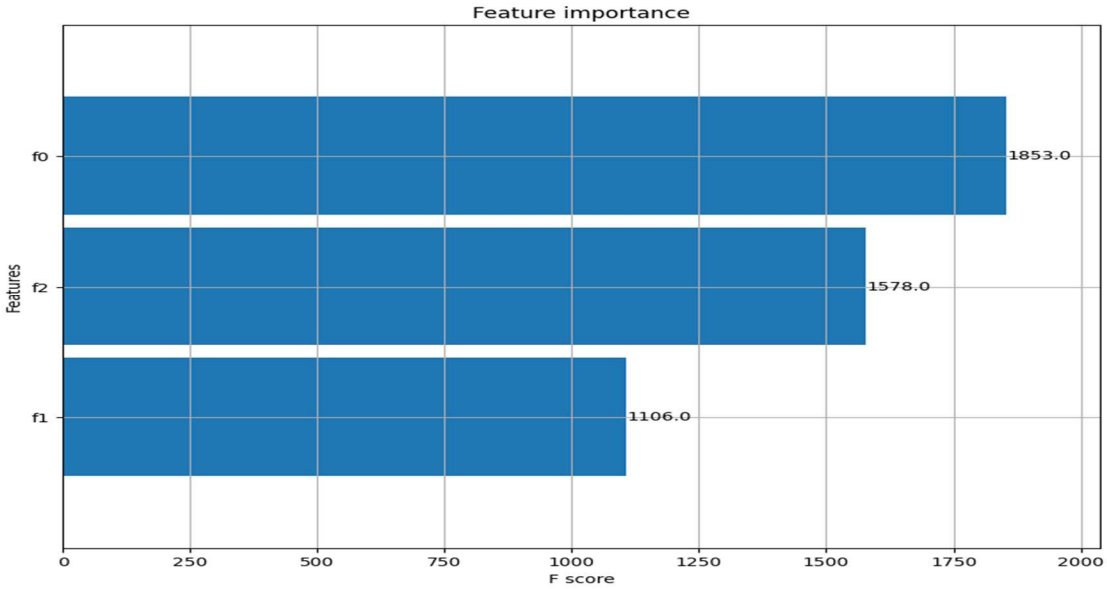
EMPIRICAL RESULTS

One unique feature of XGBoost (Extreme Gradient Boosting) is its ability to calculate feature importance. Feature importance helps one estimate how much each feature of the data contributed to the model's prediction. After performing feature importance tests, we can determine which features have the most significant impact on our model's decision-making. By removing features that have a low impact on the model's prediction and focusing on improving the more significant ones, one can significantly enhance the model's performance (Rathi, 2020).

For example, the result below displays the plot of Japan's feature importance,[ii] as shown in Figure 1. This bar chart displays *feature importance scores* based on the F-score metric for an XGBoost model. It highlights how often a feature is used to split the data across all boosting rounds.

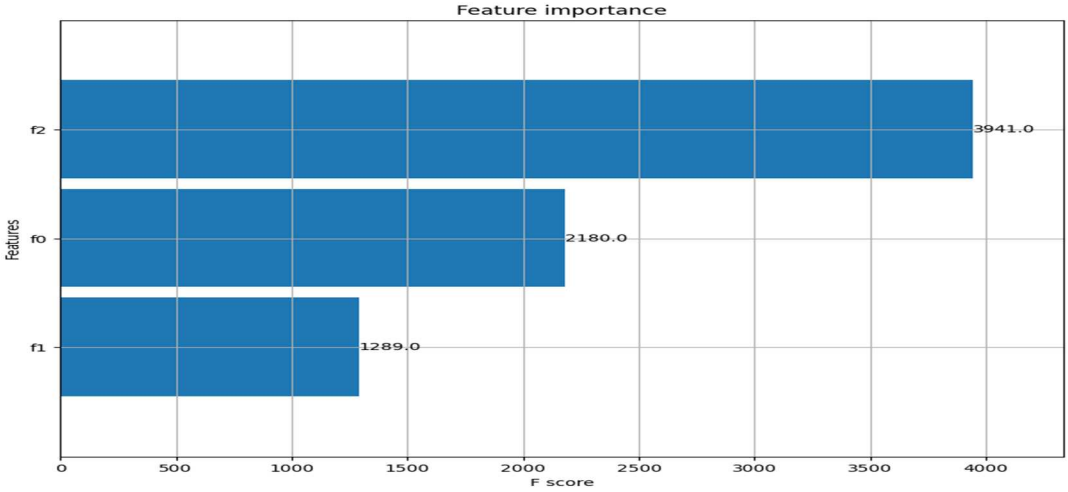
The model has only three features: GDP (f0), Unemployment Rate (f1), and Interest Rate (f2). Since this is a macroeconomic model, we assumed that these three factors can influence inflation (f4). The result indicated that GDP (f0) is the most significant factor affecting the inflation forecast.

Figure 1: Feature Importance for Japan



However, if we look at Argentina's feature importance in Figure 2, the model has only three features: GDP (f0), Unemployment Rate (f1), and Interest Rate (f2). Since this is a macroeconomic model, we assumed that these three factors can influence inflation (f4). Figure 2 revealed that the Interest Rate (f2) is the most significant factor affecting the inflation forecast for Argentina.

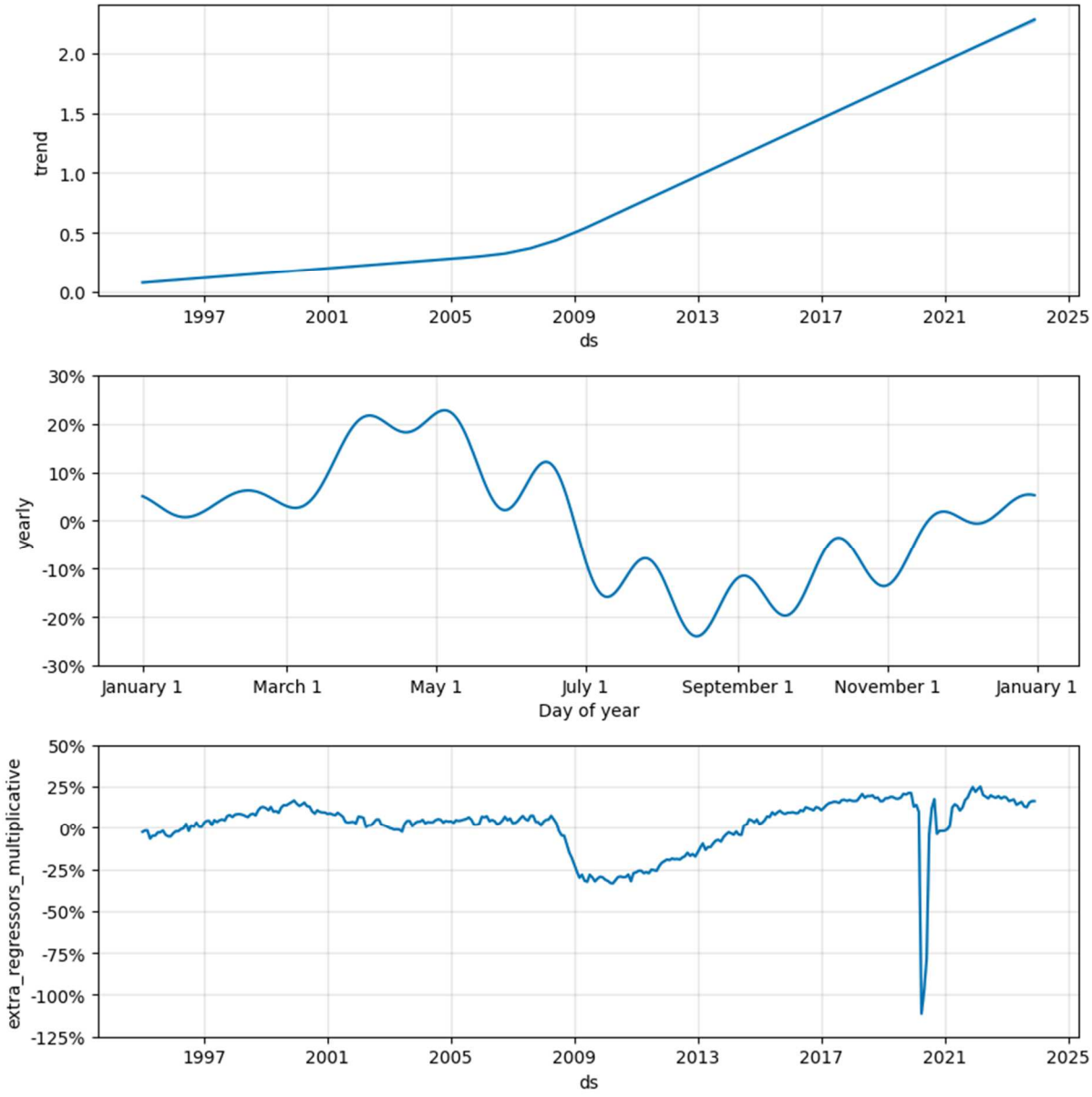
Figure 2: Feature Importance for Argentina



Analysis of different economies reveals varying importance of different features. Developed economies rely on economic performance indicators, such as GDP, whereas less developed economies, like Argentina, consider interest rates as a good predictor of inflation forecasts. This predictor may help explain why Argentina is more dependent on external debt compared to Japan.

To understand Prophet's decomposition and parameter analysis, we used the US sample data to explain how it works. The idea behind Prophet is to model time series data as a combination of trend, seasonality, and noise components. By decomposing the data into these components, the algorithm can generate accurate forecasts that capture the underlying patterns in the data. For example, the result of the decomposition for the United States data is shown in Figure 3.

Figure 3: Decomposition of the United States Data in Prophet



Analysis of Model Coefficients and Parameter Behavior

Before we analyze the results from the models, we look at the regressor coefficients and see how it relates to inflation. These results are presented in Table 1

Table 1: Regression Coefficients from FB Prophet Model

	Regressor	Mode	Center	Coefficient Lower	Coefficient	Coefficient Upper
0	USUNEM	Multiplicative	5.787	-.1103	-.1103	-.1103
1	USGDP	Multiplicative	2.470	.0090	.0090	.0090
2	USINT	Multiplicative	2.700	-.0507	-.0507	-.0507

The results of the coefficients indicated the impact of the regressor to the subject involved which is inflation. For example, in the case of GDP, the greater the GDP, the higher is the inflation for the United States.

For the parameter analysis we would focus on the seasonality prior scale and changepoint prior scale. The parameter to be tuned is the seasonality prior scale, which controls the flexibility of seasonality. A large value allows for significant seasonal fluctuation, while a low value shrinks the influence of the seasonality. We used the default value of ten. The data shows an increasing trend over the sample time period. Inflation in the US typically spikes in April and May. It eventually slows down in July and August, and then slowly increases thereafter.

Prophet detects changepoints by first specifying many changepoints at which the rate is allowed to change. It then applies a sparse prior to the magnitudes of the rate changes (equivalent to L1 regularization), as seen from the sharp decline in its regressors, particularly, which essentially means that Prophet will use as few of them as possible. In the data for the United States, we saw the impact of the COVID-19 pandemic, as seen from the sharp decline in its regressors, particularly GDP.

Creating an accurate and effective machine learning model to solve a problem is a key objective of any AI project. To optimize the model, we need to tune its parameters and hyperparameters, a crucial step in the process. This requires setting up key metrics and defining a model evaluation procedure. After implementing the changes and conducting an evaluation, we determined if the performance of the machine learning model has improved and whether we should use the updated version (Logunova, 2023).

In machine learning, parameters are the model's internal coefficients or weights that are learned and adjusted during the training process. They are estimated by fitting the training data to the model and are subsequently used to make predictions on new data. Hyperparameter tuning is the process of selecting the optimal values for the hyperparameters of a machine learning algorithm or model. It is a crucial step in developing machine learning models. Hyperparameter tuning involves finding the optimal combination of hyperparameter values that maximizes the evaluation metric. The goal is to find the set of hyperparameters that enables the machine learning model to perform best on unseen data and not generalize well to new examples. By carefully adjusting these hyperparameters, we aimed to enhance the model's ability to make accurate predictions and improve its overall performance (Logunova, 2023)

Forecastability Analysis Using Back Testing Experiments

To calculate a reliable estimation of the expected level of forecasting accuracy in terms of percentage error, an expanding window back testing strategy is implemented. Past conditions are simulated by setting the present moment anywhere in the past, building a model using historical data to forecast future inflation, and seeing how accurately it would have predicted actual data.

One can assume that, with sufficient repetitions of simulating past conditions, a highly reliable estimate of the expected level of forecastability can be calculated without waiting for a new event to occur, allowing for comparison with previously generated forecasts and concluding. To take into account yearly seasonality effects, it is recommended to carry out at least 12 repetitions of back testing experiment with one month step size should be done, which is the reason why the limit for the minimum length of observation horizon is set to 39 months (i.e. at least 24 months of historical data is required for reliable estimation of trend and/or seasonal effects, 12 months for repeated back testing experiments and three months of data to measure accuracy of quarterly forecasts).

In this research, the back testing experiment is repeated 12 times with a one-month step size for items selected from the country portfolio. At each step, a Prophet model is fitted to the historical data (i.e., the observation horizon), and monthly sales forecasts for the next three months (i.e., the forecasting horizon) are compared with the observed sales for the same period to calculate the percentage error (PE) for both monthly and quarterly forecasts. Calculated as the mean squared error (MSE) over repeated back testing experiments, which is used to quantify the expected level of forecasting reliability.

Comparative Performance Analysis of Forecasting Models

The summary of the metrics of each machine learning model is presented in Table 2:

Table 2: Evaluation Metrics of Accuracy of Each Forecasting Method

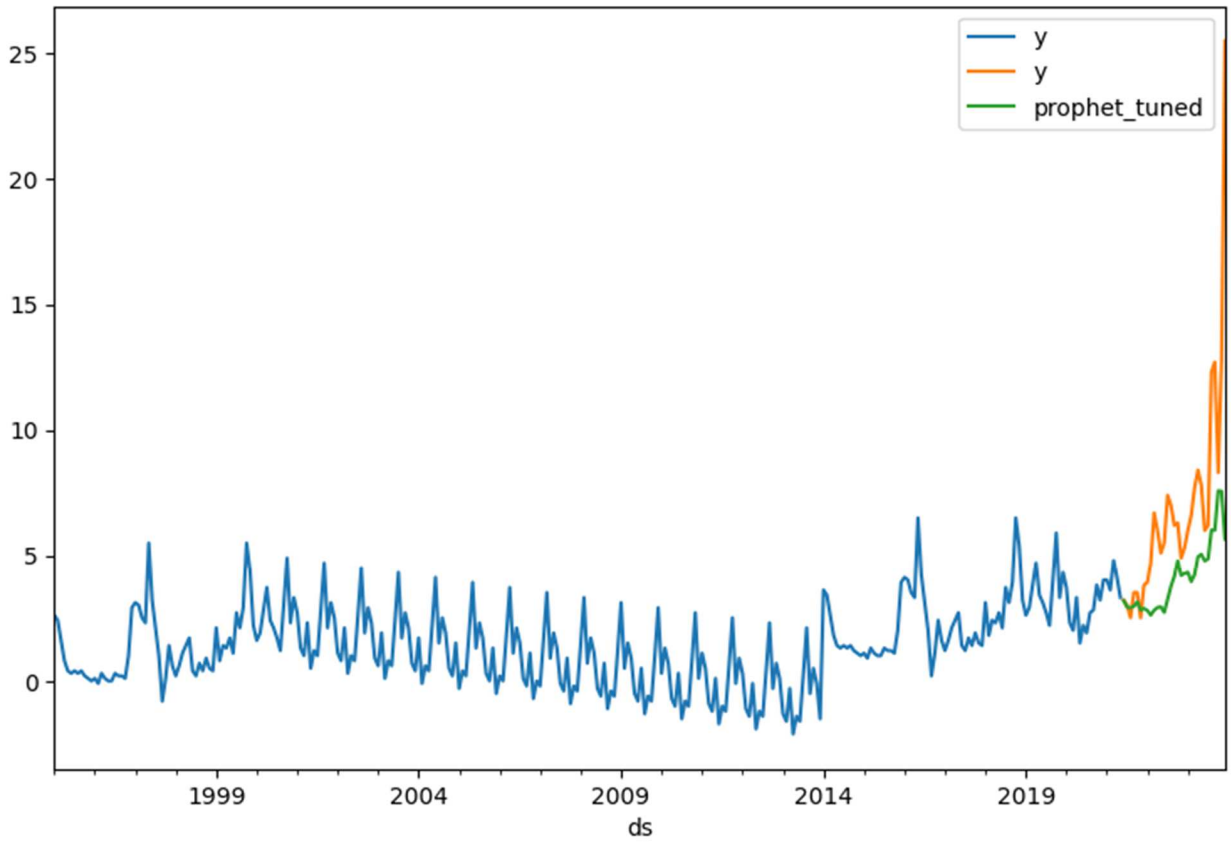
	USA	**JAPAN**	**GERMANY**	**CHINA**	**UK**	**ARGENTINA**
XGBoost:						
MSE	0	0.01	0.03	0.08	0.03	0.1
RMSE	.04	0.1	0.016	0.29	0.16	0.31
FBProphet:						
MSE	4.00	0.00	0.00	13.00	46.00	20.00
RMSE	4.00	0.00	1.00	4.00	7.00	4.00
FBProphet with CV*						
MSE	1.00	0.00	0.00	8.00	0.00	13.00
RMSE	1.00	0.00	0.00	3.00	0.00	4.00
FBProphet Tuned						
MSE	18.00	0.00	0.00	14.00	46.00	20.00
RMSE	4.00	0.00	1.00	4.00	7.00	5.00

Notes: *CV is cross-validation

The results indicated that MSE and RMSE using XGBoost are far superior to those of Prophet. Except for Japan, the MSE and RMSE values are greater than one using Prophet in most countries. However, XGBoost has a value of less than one. This suggests that using XGBoost is more efficient for forecasting than Prophet.

The result of its efficiency metrics indicates an RMSE (root mean squared error) of 0.04. However, if we look at Prophet's forecast, the US forecast is far below efficient. We can see it from the MSE/RMSE; the numbers are higher compared to the XGBoost models. Even with hypertuning under Prophet, the US forecast did not perform as well as the XGBoost model. Of the five selected countries, Japan fared the worst despite having hypertuned Prophet. Countries that performed well with hypertuning are China and Argentina. With the hypertuned Prophet model, so far, it has been the best model forecasted for Argentina. For example, an illustration of the result of the forecast with hypertuned parameters Prophet for Argentina, is given in Figure 4:

Figure 4: Prophet Model for Argentina via the Hypertuning of Parameters

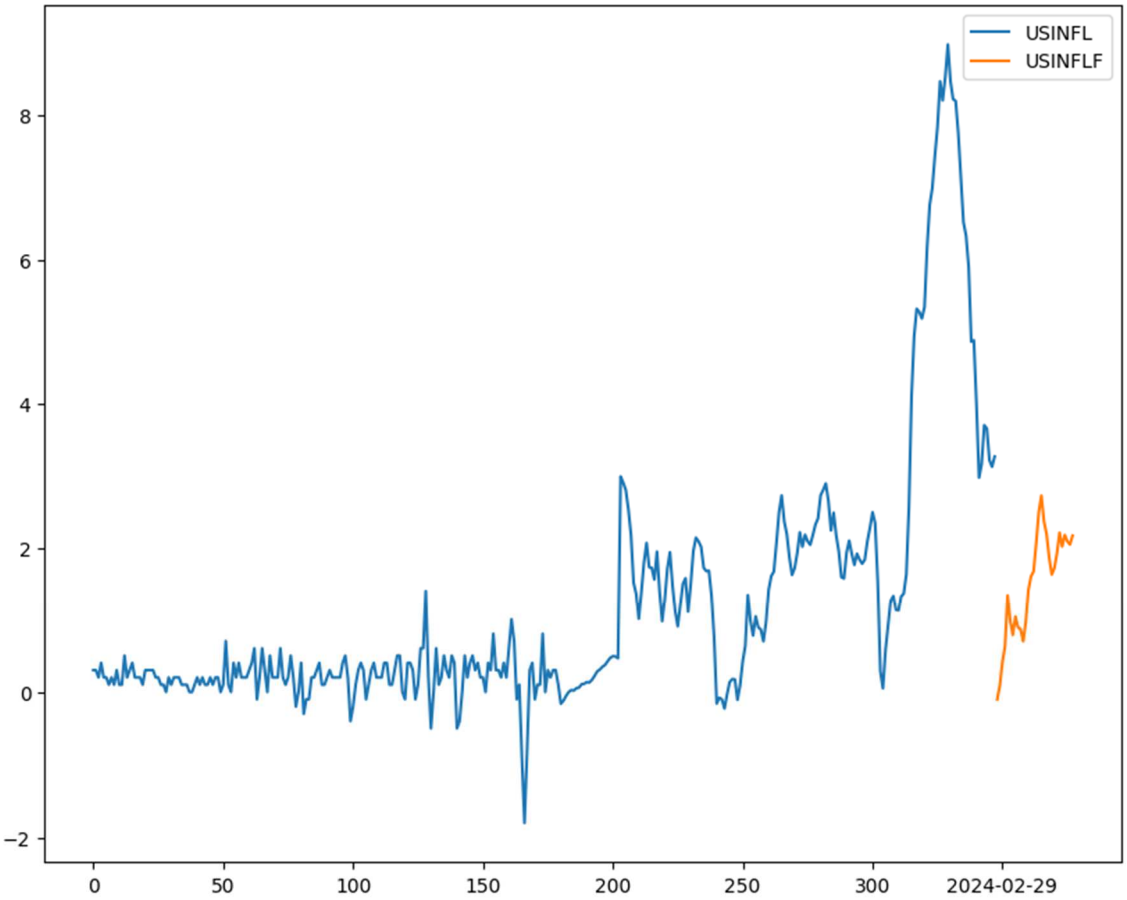


Note: the yellow line is the out of sample forecast without hypertuning while the green line is the out of sample forecast with hypertuning.

As a result of this analysis, we utilized the XGBoost model to forecast inflation beyond the sample time period, as shown in Figure 5.

The results indicated that future inflation forecasts in the United States remain around 2-3%, as shown in Figure 6. However, the trend appears to be on the rise. We should expect inflation to remain above the targeted rate, but not as high as it was compared to the previous year. Another interesting result of the future forecast is Argentina. The results indicated an average inflation rate of 5%, which is conservative. However, the trend in the forecast seems to be increasing. In both Prophet and XGBoost, the results are consistent and efficient in terms of their forecast errors. Finally, Figure 7 from 1995 to 2009, the values appear relatively stable, followed by a gradual increase with moderate variability. However, in 2019–2020, there is a pronounced surge in the orange line, likely reflecting a shock from the COVID-19 pandemic. After 2020, there is a decline, possibly representing normalization or correction following the spike.

Figure 5: Forecasting Out of Sample for the United States Using XGBoost



The Prophet model's green line closely shadows historical values with some smoothing, indicating a good fit, especially in earlier years. However, the forecast flattens slightly toward the end, possibly because Prophet emphasizes long-term trends and seasonality, and it may not fully account for abrupt anomalies without exogenous regressors. This model appears well-suited for medium- to long-term patterns but may underreact to short-term volatility. Finally, it does show the hypertuned parameter for the United States, which shows that the forecast for inflation is also 2-3% into the future. Figure 8 shows a visible cyclical

pattern repeated across the years, which Prophet is designed to model well. However, around 2019, the orange line took off sharply, showing a dramatic change. The Prophet model (green line) effectively follows this surge, suggesting it has been well-tuned with increased trend flexibility. This model for Argentina then smoothed out noise and aligned quite closely with the actual data, especially after 2019 period. The latter would be a sign of solid tuning to fit both trend and seasonality.

Figure 6: Future Inflation Forecasts for the United States

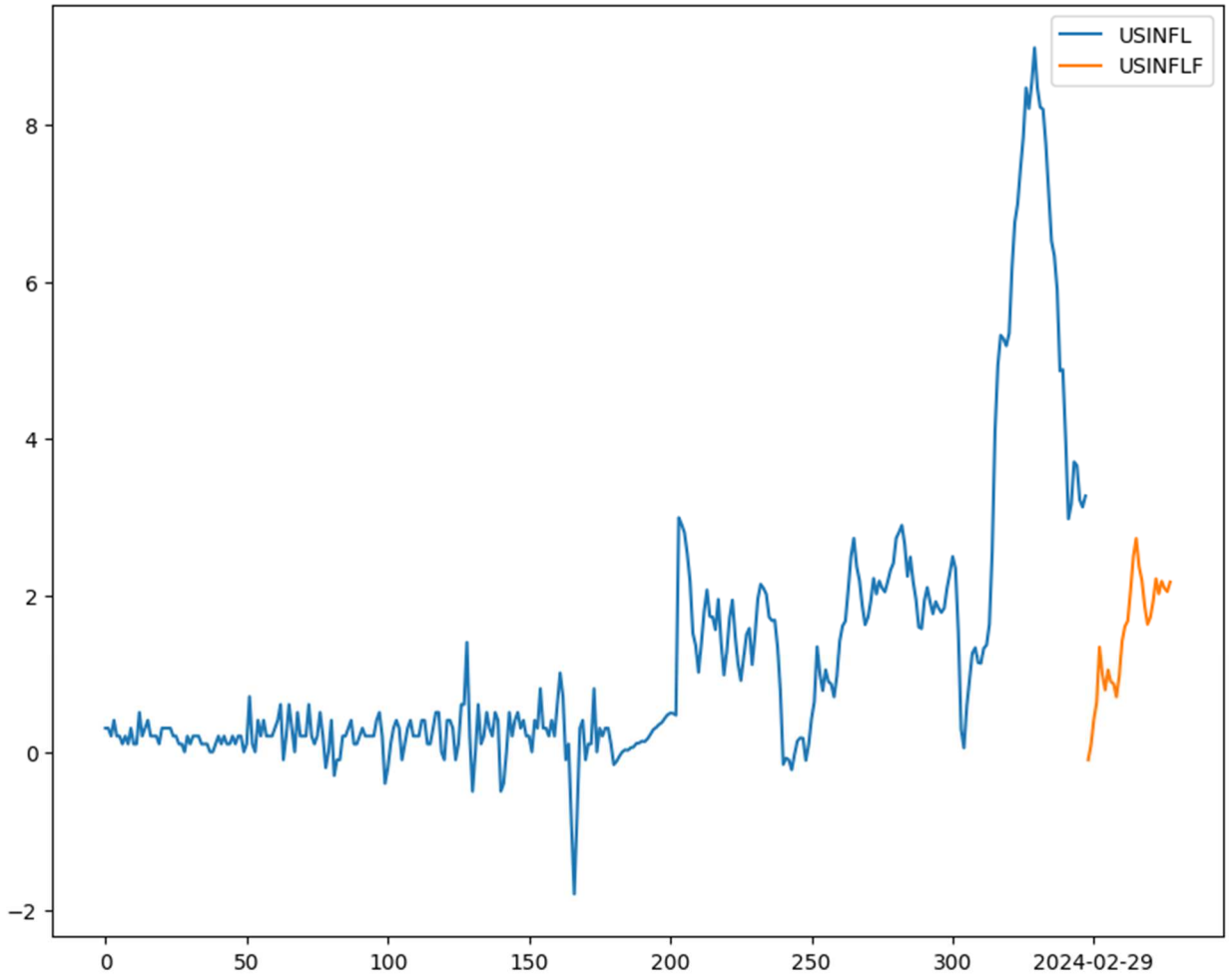


Figure 7 shows the predictions from an XGBoost model applied to a time series dataset spanning the years 2000-2024. There are some interesting results from this analysis.

1. Long-Term Trend: The model captured a general upward trajectory in “y,” suggesting a consistent increase over the 30 years. This could indicate a strong temporal component that XGBoost has identified, even though it is not inherently a time series model.

2. Model Performance: The close alignment between the predicted blue line and the actual black dots in many regions implies that the XGBoost model has fit the data reasonably well. However, there are some deviations, particularly around 2020, which is attributed to the COVID-19 pandemic.
3. Uncertainty Range: The shaded blue region reflects the confidence interval for the predictions. As we move beyond the last observed point into the forecast range (past 2023), this region widens, highlighting increasing uncertainty in the model's future predictions.

Figure 7: Analysis of Forecasts based on XGBoost Model

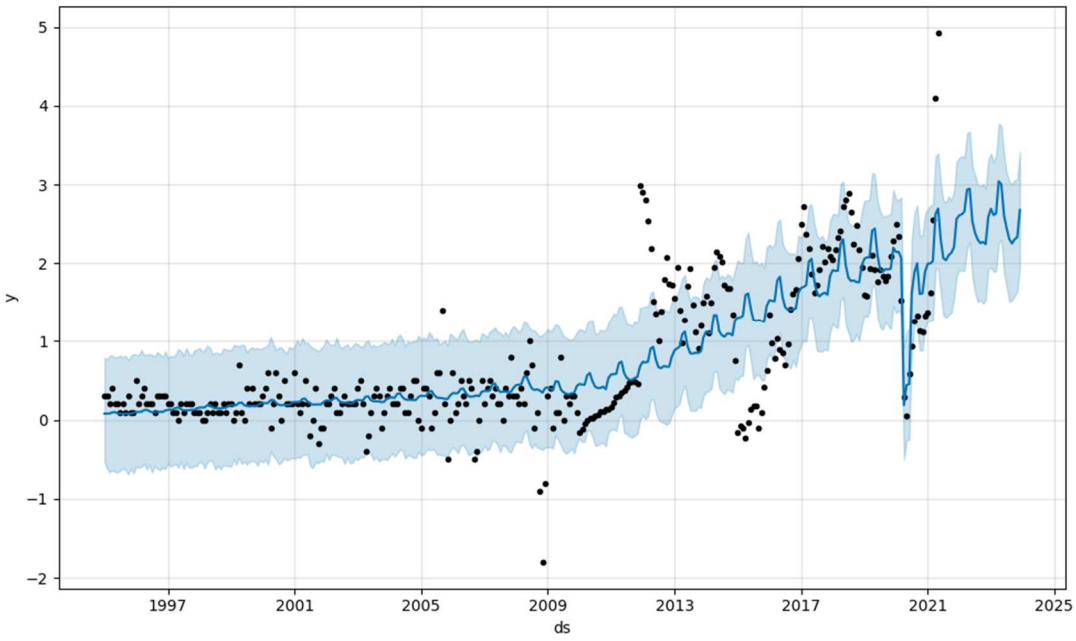


Figure 7: Prophet Hypertuned Forecast for the United States:

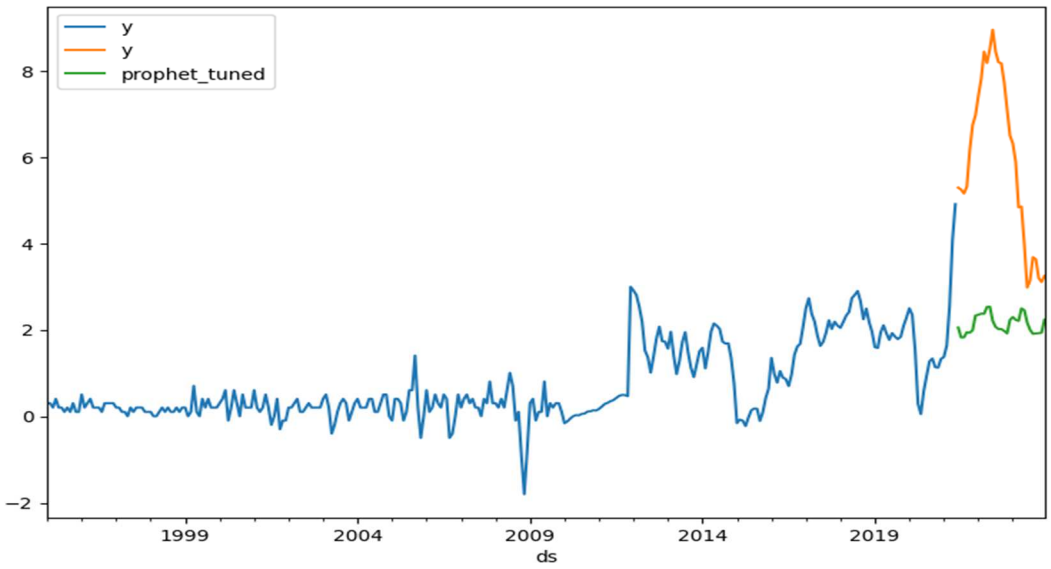
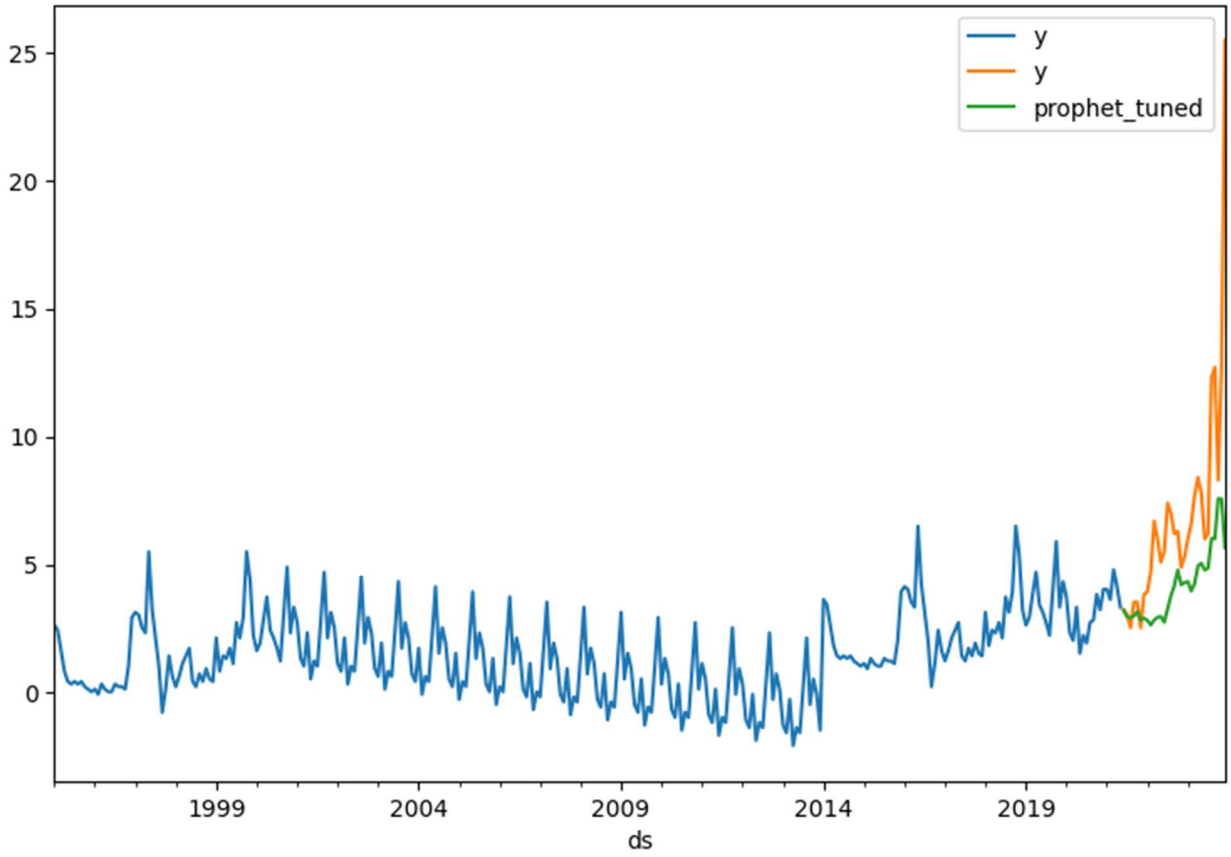


Figure 8: Prophet Hypertuned Forecast for Argentina:



CONCLUSION AND FINAL THOUGHTS

Prophet provided valuable insight into the historical inflation data by decomposing it into three key components: trend, seasonality, and noise. This decomposition allowed for a clearer understanding of the underlying patterns within the data. Notably, one interesting observation from this breakdown was the seasonal rise in inflation during specific months of the year, which highlighted the model's strength in revealing temporal dynamics. However, while Prophet was helpful for exploratory analysis and pattern recognition, its forecasting performance lagged behind that of XGBoost in terms of accuracy.

When comparing the two models in forecasting U.S. inflation, both predicted a gradual and steady trend in inflation for future out-of-sample periods. Nevertheless, XGBoost significantly outperformed Prophet, as indicated by a much lower Root Mean Square Error (RMSE). This superior performance was further enhanced by careful hyperparameter tuning, which contributed to the reduction in RMSE and improved the overall forecast accuracy of the XGBoost model.

Beyond these metrics, the flexibility of XGBoost to handle complex interactions and non-linear relationships among features gave it a notable advantage, especially when multiple, interdependent economic indicators influenced historical inflation patterns. Additionally, XGBoost's ability to incorporate a broader set of engineered features—such as lag variables, macroeconomic factors, and calendar effects—allowed it to capture nuances that Prophet, with its default additive structure, might have overlooked.

REFERENCES

Almosova, A., & Andresen, N. (2023). Nonlinear inflation forecasting with recurrent neural networks. *Journal of Forecasting*, *42*(2): 240–259.

Baybuza, I. (2018). Inflation forecasting using machine learning methods. *Russian Journal of Money and Finance*, *77*(4): 42–59.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., ... & Zhou, T. (2015). Xgboost: Extreme gradient boosting. *R package version 0.4-2*, *1*(4), 1-4.

Chen, T., & Guestrin, C. (2016). *XGBoost: A scalable tree boosting system*. In B. Krishnapuram, M. Shah, A. J. Smola, C. Aggarwal, D. Shen, & R. Rastogi (Eds.), *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). Association for Computing Machinery.

Choudhary, M. A., & Haider, A. (2012). Neural network models for inflation forecasting: an appraisal. *Applied Economics*, *44*(20): 2631–2635.

Garcia, M. G., Medeiros, M. C., & Vasconcelos, G. F. (2017). Real-time inflation forecasting with high-dimensional models: The case of Brazil. *International Journal of Forecasting*, *33*(3): 679–693.

Faust, J., & Wright, J. H. (2013). Forecasting inflation. In *Handbook of Economic Forecasting* (Vol. 2, pp. 2-56). Elsevier.

Fuhrer, C. (2010). Inflation persistence. Handbook of monetary economics. *Handbook of Monetary Economics, edition*, *1*: 423–486.

Gonzalez, S. (2000). Neural networks for macroeconomic forecasting: A complementary approach to linear regression models. Available at: Working Paper

Hebb, D. O. (1949). *The organization of behavior: A neuropsychological theory.* New York: Wiley.

Khare, P. (2023), Understanding FB Prophet: A Time Series Forecasting Algorithm Medium.(Illumination) May 13, 2023.

Logunova, I. (2023) Hyperparameter Tuning and Evaluation of ML Models https://serokell.io/blog/hyperparametertuning-in-ml. July 30, 2023.

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics, 5*(4): 115–133.

Medeiros, M. C., Vasconcelos, G. F. R., Veiga, Á., & Zilberman, E. **(2021).** Forecasting inflation in a data-rich environment: The benefits of machine learning methods. *Journal of Business & Economic Statistics*, *39*(1): 98–119.

Menculini, L., Marini, A., Proietti, M., Garinei, A., Bozza, A., Moretti, C., & Marconi, M. (2021). Comparing prophet and deep learning to ARIMA in forecasting wholesale food prices. *Forecasting*, *3*(3): 644-662.

Moshiri, S., & Cameron, N. (2000). Econometrics versus ANN models in forecasting inflation. *Journal of Forecasting*, *19*(3): 201–217.

Nakamura, E. (2005). Inflation forecasting using a neural network. *Economics Letters*, *86*(3): 373–378.

Özgür, Ö., & Akkoç, U. (2022). Inflation forecasting in an emerging economy: selecting variables with machine learning algorithms. *International Journal of Emerging Markets*, *17*(8): 1889-1908.

Purnawansyah, P., Haviluddin, H., Setyadi, H. J., Wong, K., & Alfred, R. (2019). An inflation rate prediction based on backpropagation neural network algorithm. *International Journal of Artificial Intelligence Research*: *3*(2).

Rathi, P. (2020). A novel approach to feature importance—Shapley Additive Explanations. *Towards Data Science*, *2*.

Rodríguez-Vargas, A. (2020). Forecasting Costa Rican inflation with machine learning methods. *Latin American Journal of Central Banking*, *1*(1-4): 100012.

Smalter Hall, A., & Cook, T. R. (2017). Macroeconomic indicator forecasting with deep neural networks. *Federal Reserve Bank of Kansas City Working Paper*, 17-11.

Stock, J. H., & Watson, M. W. (1998)**.** Asymptotically median unbiased estimation of coefficient variance in a time-varying parameter model. *Journal of the American Statistical Association, 93*(441): 349–358.

Stock, J. H., & Watson, M. W. (2007). Why has US inflation become harder to forecast?. *Journal of Money, Credit and Banking*: *39*, 3–33.

Szafranek, K., & Polski, N. B. (2017). *Bagged artificial neural networks in forecasting (low) inflation: an extensive comparison with current modelling frameworks* (No. 262). NBP Working Paper.

Tkacz, G., & Hu, S. (1999). *Forecasting GDP growth using artificial neural networks* (No. 1999-3). Bank of Canada.

Wolters, M. H., & Tillmann, P. (2015). The changing dynamics of US inflation persistence: A quantile regression approach. *Studies in Nonlinear Dynamics & Econometrics*, *19*(2): 161–182.

Zahara, S., & Ilmiddaviq, M. B. (2020). Consumer price index prediction using Long Short-Term Memory (LSTM) based cloud computing. In *Journal of Physics: Conference Series* 1456, (1): 012022.

Zunic, E., Korjenic, K., Hodzic, K., & Donko, D. (2020). Application of Facebook's Prophet algorithm for successful sales forecasting based on real-world data. *arXiv preprint arXiv:2005.07575*.

[i] Monte Carlo uses randomness to estimate values. A Markov chain is a sequence of possible events where the probability of each event depends only on the state attained in the previous event, i.e., it has no memory. Then the MCMC uses a Markov chain to generate samples from a probability distribution.

[ii] Feature importance in XGBoost refers to scores that indicate how useful each feature was in building the decision trees within the model. It is calculated by evaluating how often a feature is used in splits, how much it improves model performance when used, or how many samples it covers. In summary, these scores help identify which features are most influential in making predictions.

Causal Diagrams and Structural Debriefing in Macroeconomics: A Literacy-Targeted Pilot Study

Oleg V. Pavlov[1], Natalia V. Smirnova[2], Elena Smirnova[3]

ABSTRACT

A literacy-targeted approach to economics instruction emphasizes depth over breadth, drawing on cognitive science principles that highlight the role of schema construction in learning complex material. Building on this perspective, we designed and implemented instructional activities that integrate *causal diagrams* with *structural debriefing*, intended to support schema development while promoting deeper understanding of macroeconomic concepts beyond traditional lecture-based methods. Results from the pilot implementation suggest that diagram-based debriefing activities were associated with improvements in students' ability to identify causal relationships. The innovation of this paper lies in the instructor-ready materials provided, making it valuable to practitioners seeking actionable teaching tools.

INTRODUCTION

Economic instruction often requires students to process complex models and relationships, yet traditional lecture-based methods can overwhelm learners and limit deep understanding. Cognitive science suggests that comprehension and retention depend on the construction and refinement of schemas**,** which are networks of interconnected concepts that organize knowledge. Literacy-targeted teaching builds on this insight by emphasizing deliberate practice and in-depth exploration of fewer topics to foster schema development.

This paper contributes to the existing literature by providing a novel literacy-targeted approach to economics instruction. We introduce a new method that integrates causal diagrams with structural debriefing activities. Causal diagrams visually represent cause-and-effect relationships among economic variables, offering students a tool for organizing material into coherent schemas. Structural debriefing guides learners through analyzing these diagrams, reinforcing understanding of feedback loops and multiplier effects.

Using a design-based approach, we implemented two pilot activities in an intermediate macroeconomics course: one focused on the government-purchases multiplier and the other on the tax

[1] Social Science and Policy Studies, Worcester Polytechnic Institute, opavlov@wpi.edu

[2] Department of Economics, University of Connecticut, natalia.smirnova@uconn.edu

[3] Department of Management, Marketing and Finance, School of Business, State University of New York at Old Westbury, smirnovae@oldwestbury.edu

multiplier. We report findings on student performance and feedback, demonstrating the potential of diagram-based schema construction to enhance economic literacy.

The innovation and primary strength of this paper lies in its creative pedagogical design, the synthesis of system dynamics with structural debriefing, and the instructor-ready materials provided, making it highly valuable to practitioners looking for actionable teaching tools. This paper's primary contribution is a pedagogical framework and an accompanying instructional activity, rather than an empirical study establishing the causal efficacy of causal diagrams on learning outcomes. The quantitative and qualitative results we report are intended to illustrate the framework's use in practice rather than to serve as its validation.

The remainder of this paper is organized as follows. The next section reviews innovations in economics instruction, specifically a literacy-targeted approach to enhancing learners' comprehension of the material. We then introduce the concept of causal diagrams, outline our research method, and explain the design process used to develop the instructional activities. Subsequent sections describe the implementation of structural debriefing activities and present findings from classroom observations. The paper concludes with a discussion of implications for economics education and suggestions for future research.

COGNITIVE FOUNDATIONS OF LITERACY-TARGETED TEACHING

Economics education has evolved through a range of pedagogical innovations aimed at improving student engagement, conceptual understanding, and literacy-targeted instruction. The *literacy-targeted approach* to economics instruction emphasizes depth over breadth, focusing on fewer topics to promote conceptual clarity and retention (Goffe and Wolla, 2024). This approach draws on principles from cognitive and learning sciences, including schema theory, constructivism, and deliberate practice.

Human cognitive architecture lies at the core of literacy-targeted teaching. Cognitive learning theory uses that architecture to generate instructional hypotheses relevant to teaching (Sweller, 1999; Sweller, 2025). A student receiving and processing new material is constrained by working memory, which has limited capacity and duration. Long-term memory, by contrast, has no known capacity limit: it stores domain-specific knowledge structures, or schemas, networks of interconnected ideas, concepts, and procedures that help learners organize, interpret, and retrieve information (Goffe and Wolla, 2024: p. 157; Kalyuga et al., 2003).

Student comprehension of economic models depends on the development of these schemas, which serve as cognitive structures for integrating new knowledge. According to constructivist theory, learners actively build and refine schemas by connecting new ideas to prior knowledge (Ambrose et al., 2010). As novices progress toward expertise, their schemas become more sophisticated and interconnected.

Deliberate practice complements this process by introducing instructional activities specifically designed to strengthen schema formation (Goffe and Wolla, 2024: p. 159). Together, these principles provide a foundation for literacy-targeted teaching strategies that aim to help students process complex economic relationships through structured, schema-building exercises.

Kalyuga, Ayres, Chandler and Sweller (2003) argue that the level of skills, knowledge, and learning styles of students influence the method of instruction. For more experienced learners who rely on schemas to process complex material, a guided approach to instruction may actually be detrimental, an effect the authors term the Expertise Reversal Effect. It implies that to be efficient, the instructional activity has to be tailored to the level of the student audience. Instructional techniques effective for novices, who have not yet built the relevant schemas, may become redundant or even counterproductive for advanced students who have. After sufficient practice, schemas can operate automatically, requiring minimal working memory, and experienced learners are able to recognize a familiar pattern as a single schema rather than multiple new pieces of information (Kalyuga et al, 2003: p. 24), reducing working memory strain.

For example, diagrams integrated with text explanations benefit inexperienced learners but prove redundant for knowledgeable trainees (Kalyuga, Chandler and Sweller, 1998). Novices could not comprehend diagram-only format and needed the accompanying text. By contrast, advanced trainees ignored the redundant text and reported less mental effort when studying the diagram alone. This observation is relevant to our paper since we introduce a novel causal diagram to represent the government-purchases and tax multiplier effects.

In the following section, we describe how these principles informed the design of our instructional approach, which integrates causal diagrams and structural debriefing into macroeconomics teaching.

METHOD

Research in cognitive science and education shows that students often fail to recognize complex causal relationships when learning new material (Perkins and Grotzer, 2005; Grotzer, 2012; Sterman, 2010; Sweller, 2025). We extend this observation to economics education, where understanding causal complexity is critical for building coherent economic schemas. Introductory and intermediate economics courses present students with numerous models and concepts, yet many struggle to connect these ideas into an integrated framework (Goffe and Wolla, 2024).

To address this challenge, we designed and implemented a set of pedagogical activities that combine causal diagrams with structural debriefing, helping students identify key variables and causal connections within the models typically covered in a macroeconomics textbook and fostering schema construction. We integrate these diagrams into the curriculum through structural debriefing activities, adapted to explain causal relationships discussed in macroeconomic texts, visualizing economic variables and their interrelationships as graphical diagrams. Each activity can be incorporated into traditional lecture-based instruction: causal diagrams supplement textual explanations and conventional graphs, such as supply-and-demand curves, by providing an additional visual representation of economic relationships.

To design and refine these activities, we employed a design-based research approach (Collins et al., 2004; McKenney and Reeves, 2012). This iterative process allowed us to improve the activities based on classroom observations, student performance, and feedback. Our aim was to create lesson plans and materials that economics instructors could use without prior expertise in causal diagrams.

The following subsections provide details on causal diagrams, structural debriefing, and the study design.

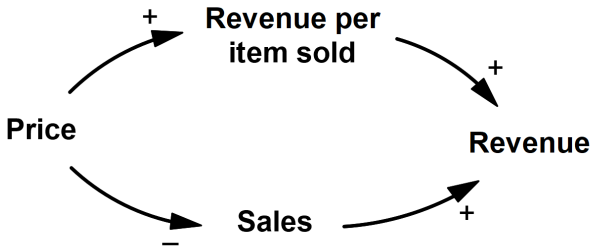
Causal diagrams

Causal diagrams are simple yet powerful tools for representing cause-and-effect relationships. Commonly used in systems thinking, system dynamics (Maani and Cavana, 2007; Sterman, 2000), and artificial intelligence (Pearl and Mackenzie, 2018), these diagrams depict variables as nodes and causal relationships as arrows. A positive arrow indicates that the cause and effect move in the same direction, while a negative arrow shows an inverse relationship. The same conventions underline causal diagrams used in feedback-based economic modeling, where reinforcing and balancing feedback loops help explain the dynamic behavior of economic systems over time (Cavana et al., 2021).

Figure 1 illustrates how this notation can represent the two effects of price on revenue:

- Positive effect: Revenue per item sold increases with price, represented by a positive arrow. Higher marginal revenue contributes to overall revenue, forming a causal chain of two positive arrows that capture the combined positive effect of higher prices on revenue.
- Negative effect: A higher price reduces sales, shown by a negative arrow. Because sales and revenue move in the same direction, the arrow between them is positive. This sequence of a negative arrow followed by a positive arrow forms a negative causal chain, capturing the negative effect of higher prices on revenue due to declining sales.

Figure 1: Example of a causal diagram illustrating the effect of price on revenue. The diagram shows two causal chains: the positive effect (upper chain) and the negative effect (via sales).



Prior research demonstrates the effectiveness of causal diagrams in economics education. Jägerskog and colleagues (2019; 2021a; 2021b) found that these diagrams help students achieve a deeper understanding of economic material compared to traditional supply-and-demand graphs alone.

Structural debriefing

Debriefing generally refers to reviewing and analyzing an event**,** such as a business simulation, military operation, or surgery**,** with participants who experienced it (Crookall, 2010; Lederman, 1992). Debriefing activities often include discussion, journaling, analysis, personalization, and generalization of the experience (Lederman, 1992).

Structural debriefing is a specialized form of debriefing in which participants identify key variables, discuss causal relationships among them, and connect system behavior to its underlying causal structure (Pavlov et al., 2015). This method originated in the context of instructional simulations, where it was used to help participants understand a simulation's behavior in terms of its underlying causal structure, rather than treating the simulation as a black box (Pavlov et al., 2015). Structural debriefing has been shown to improve performance in instructional management simulations (Capelo and Silva, 2020; Capelo et al., 2021; Capelo et al., 2024; Qudrat-Ullah, 2020) and educational video games (Kim and Pavlov, 2019; Pavlov et al., 2019).

Study design

We adopted a design-based research methodology (Collins et al., 2004; McKenney and Reeves, 2012) to conduct formative research aimed at validating and refining the activities. This approach allowed us to adapt general principles to specific classroom contexts and reflect on emerging findings. After implementing each activity, we analyzed classroom observations, student performance, and feedback, then iteratively revised the materials.

Our research question was: How can instructors utilize causal diagrams and structural debriefing activities for literacy-targeted instruction in a macroeconomics classroom?

In early discussions, we identified design principles for lesson plans that could be used by instructors with no prior experience in causal diagrams. Key considerations included limited classroom time and instructors' unfamiliarity with the approach. We established the following design principles:

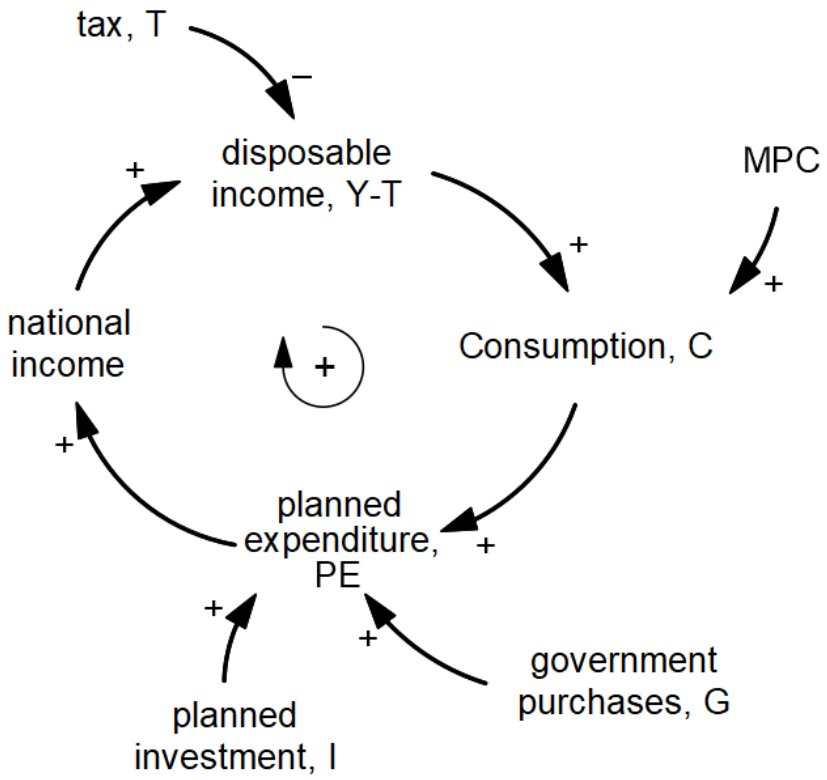
1. **Develop self-contained activities for macroeconomic topics:** We based our activities on Mankiw's Macroeconomics (2025), a widely used intermediate-level textbook. Although the textbook builds concepts across chapters, we limited each causal diagram to a single topic for simplicity. Activities were designed as standalone units, giving instructors flexibility in selecting which to include.
2. **Simplify tasks for students:** While structural debriefing can begin with a blank page, asking students to identify variables and construct diagrams from scratch could be overwhelming. To reduce cognitive load, each activity includes a causal skeleton, which is a diagram with variables and undirected edges, providing a starting point for students.
3. **Provide instructors with ready-to-use materials:** To minimize adoption barriers, we developed detailed lesson plans, slides, and assignments for each activity.

After reviewing the syllabus for an Intermediate Macroeconomics course taught by one of the authors, we prepared debriefing activities for the government-purchases multiplier and the tax multiplier.

Activity 1 – Government-purchases multiplier: Figure 2 presents a causal diagram designed to illustrate the government-purchases multiplier discussed in Mankiw's (2025) chapter on Aggregate

Demand. The diagram includes eight key variables and eight causal connections, with only one negative arrow representing the relationship between lump-sum tax (T) and disposable income (Y – T). The circular causal connections form a positive feedback loop, indicated by a circular arrow with a plus sign in the center. This feedback mechanism, explicitly described in Mankiw (2025), explains how the government-purchases multiplier effect occurs. The diagram helps students visualize this feedback process and understand its role in amplifying the initial change in government spending.

Figure 2: The causal diagram for the activities on the multiplier effects.



Activity 2 - Tax multiplier: The same causal diagram in Figure 2 is used for the tax multiplier activity. In this activity, students explore how changes in tax (T) propagate through the system and affect other variables. By tracing these causal pathways, students gain insight into the mechanisms underlying the tax multiplier effect. All activity materials distributed to students, including worksheets and instructions, are provided in the appendices at the end of this article.

IMPLEMENTATION IN THE CURRICULUM

This section describes the implementation of structural debriefing activities during an Intermediate Macroeconomics course at a large public research university. The course used Mankiw's Macroeconomics (2025) as the primary textbook. Activities 1 and 2 were combined and implemented during a single 75-minute class session. Prior to these activities, the relevant textbook chapters were covered through traditional instruction, including lectures and homework assignments.

Sample description

The course was a required "Quantitative Intermediate Macroeconomics" section, carrying a "Q" designation for courses that emphasize the quantitative assessment of macroeconomic variables such as GDP, business cycles, inflation, unemployment, and growth, and the empirical evaluation of fiscal and monetary policy. Beyond exams and homework, the course includes a semester-long empirical research project with peer review and an oral presentation component. Students who elect to enroll in a Q-designated section are therefore likely to be more quantitatively inclined, on average, than students in a standard intermediate macroeconomics course.

The class had 13 students in total: 5 master's students and 8 undergraduates. This course was the first required course in the accelerated master's program, so graduate students had only a marginal advantage in preparedness, since much of the relevant material had to be recalled from undergraduate coursework completed, in some cases, several years earlier. As a result, prior preparation was broadly comparable across the two groups. The combination of undergraduate and master's students in a single section reflects an administrative scheduling matter rather than any deliberate sampling design for this study.

Because participants were drawn from a single, quantitatively self-selected section at one university, and because the sample was small (n = 13), the findings reported below should be interpreted as evidence from a specific instructional context rather than as broadly generalizable to intermediate macroeconomics students. We return to this limitation, along with the absence of a control group, in the Limitations and Future Research section.

Treatment of the multiplier in other textbooks

The Mankiw "Macroeconomics" (2025) textbook was used because it was already the required text for the course in which these activities were piloted. Structural debriefing activities are built around the causal structure a given textbook presents, so we designed our causal diagram to follow Mankiw's step-by-step exposition of the multiplier, which builds the expenditure multiplier through the traditional Keynesian Cross and presents an explicit geometric series of successive consumption rounds based on the marginal propensity to consume. This structure allowed us to isolate each round of the feedback loop for students.

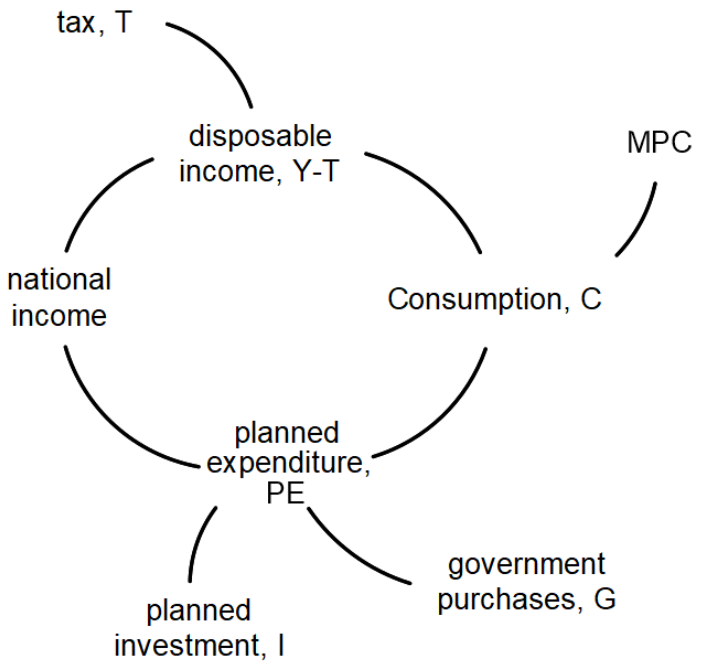
To help readers assess the broader applicability of our approach, we compared this treatment to two other popular intermediate macroeconomics textbooks. Blanchard et al. (2026) instead model investment endogenously as a function of the interest rate and output, and Abel, Bernanke, and Croushore (2024)

mute the multiplier through crowding out while de-emphasizing the standalone Keynesian Cross in favor of a classical-Keynesian synthesis. Teaching materials for courses that use either of these alternative texts would need to adapt the causal diagram in Figure 2 to reflect the different investment specifications, illustrating how a structural debriefing activity must be tailored to the causal structure of the textbook it accompanies.

Activity 1: Government-purchases multiplier

Thirteen students participated in Activity 1, working individually. Although students had previously studied the government-purchases multiplier, the activity included relevant passages from the textbook to reinforce the concepts. Each student received a hard copy of Form 1 (see Appendix A), which included a causal skeleton graph (Figure 3). Using the provided quotes, students were asked to complete the graph by adding causal directions and link polarities. They were given 10 minutes to complete this task.

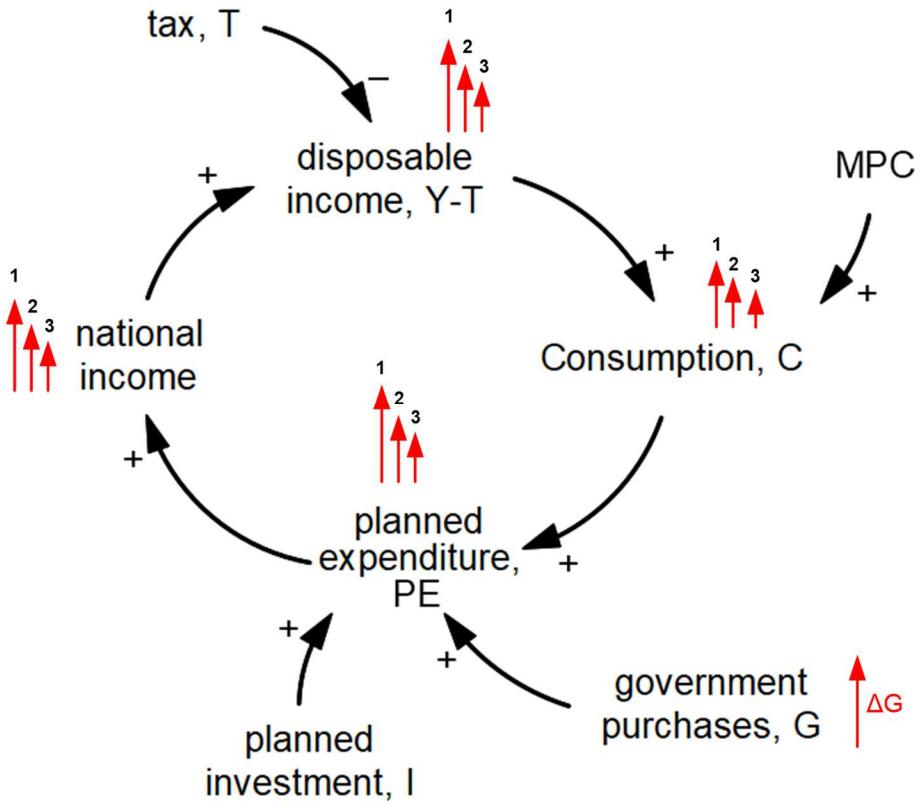
Figure 3: Causal skeleton graph provided to students during Activity 1.



After submitting their responses, students were shown the completed causal graph. The instructor then used a prepared slide deck to explain, step by step, how an initial increase in government purchases, ΔG, propagates through the feedback loop, resulting in the multiplier effect. Figure 4 illustrates this process. When government purchases increase by ΔG, planned expenditure, national income and disposable income rise by ΔG as well. However, consumption increases only by $(MPC \times \Delta G) < \Delta G$ because $MPC < 1$.

This initial change in consumption triggers a second iteration around the feedback loop, increasing planned expenditure by *MPC x ΔG.* Each subsequent iteration produces a smaller effect than the previous one. The numbers above the vertical red arrows in Figure 4 correspond to the first, second, and third iterations, while the arrow lengths indicate the relative magnitude of each effect.

Figure 4: Government-purchases multiplier effect. Numbers above the vertical red arrows indicate the first, second, and third iterations around the feedback loop.



Corroborating Mankiw's (2025, p. 280) textbook explanations, the instructor emphasized that the total effect of *ΔG* on national income is:

Initial Change in Government Purchases	= ΔG
First Change in Consumption	= MPC x ΔG
Second Change in Consumption	= MPC^2 x ΔG
Third Change in Consumption	= MPC^3 x ΔG
⋮	⋮

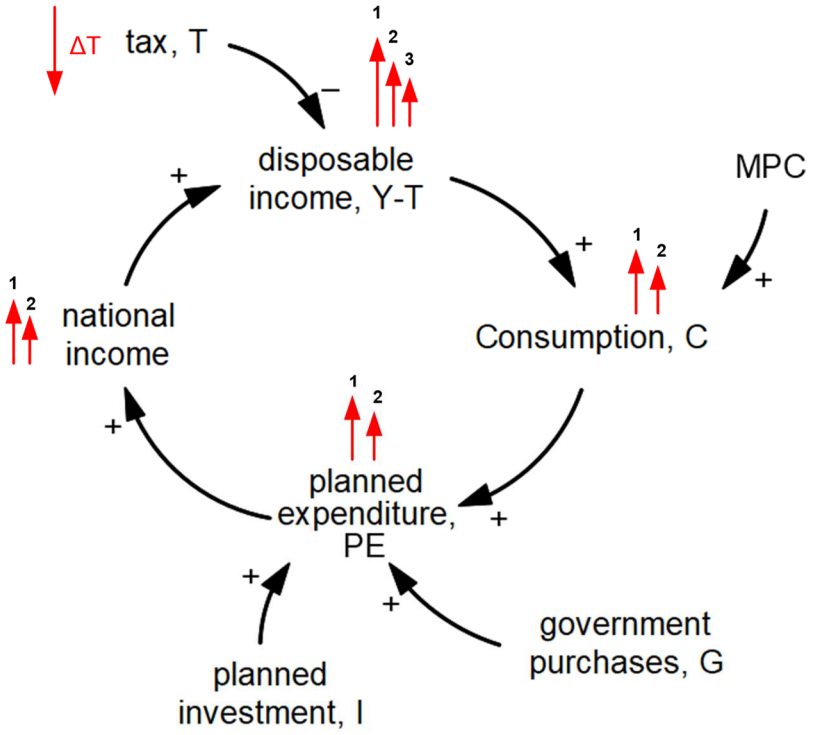
$$\Delta Y = (1 + MPC + MPC^2 + MPC^3 + \dots)\,\Delta G$$

Activity 2: Tax multiplier

Activity 2 followed a similar structure. Students received Form 2 (see Appendix B) that quoted relevant passages from Mankiw's (2025) discussion of the tax multiplier effect. The form included the same causal skeleton as in Figure 3, which students were asked to complete by adding causal directions and link polarities. This was their second time working with the same causal skeleton during the session.

Next, students were asked to explain the tax multiplier effect using the completed causal graph. This analysis built on the approach introduced in Activity 1, where the instructor demonstrated how an initial increase in government purchases produces multiple rounds of consumption changes. After submitting their worksheets, students were shown the completed causal diagram illustrating the effect of a tax decrease, ΔT (Figure 5).

Figure 5: Tax multiplier effect. Vertical red arrows indicate the first, second, and third iterations around the feedback loop.



The instructor then explained that the total effect on national income is:

Initial Change in Taxes	= - ΔT
First Change in Consumption	= MPC x ΔT
Second Change in Consumption	= MPC^2 x ΔT
Third Change in Consumption	= MPC^3 x ΔT
⋮	⋮

ΔY = (MPC + MPC^2 + MPC^3 +…) ΔT

As noted in Mankiw (2025), there is an inverse causal relationship between taxes, T, and disposable income, $Y - T$. Therefore, a tax decrease of $-\Delta T$ results in an increase in disposable income ΔT, which leads to the first change in consumption equal to $MPC \times \Delta T$.

Post-activity reflection

After completing Activities 1 and 2, students answered five questions (see Form 3 in Appendix C). Two questions asked students to compare the government-purchases and tax multipliers, focusing on similarities and differences. These comparison questions were inspired by Goffe and Wolla's (2024: 159) recommendation to use deliberate practice for schema development in economics courses, such as by asking students to analyze how income and price elasticities are similar and different.

FINDINGS

This section analyzes the implementation of the two activities. Overall, students enjoyed the exercises and were actively engaged.

Activity 1: Government-purchases multiplier

Table 1 summarizes the results of the "Government-Purchases Multiplier" exercise, in which students completed a causal skeleton by adding arrow directions and polarities. Thirteen students submitted worksheets (see Appendix A). The causal diagram had eight variables and eight edges.

Table 1: Results of completing the causal diagram for the government-purchases multiplier

Student	Correct answers		Percentage correct	
	Direction	Polarity	Direction	Polarity
1	8	6	100.00%	75.00%
2	7	0	87.50%	0.00%
3	7	6	87.50%	75.00%
4	7	0	87.50%	0.00%
5	0	3	0.00%	37.50%
6	7	7	87.50%	87.50%
7	5	8	62.50%	100.00%
8	5	8	62.50%	100.00%
9	6	8	75.00%	100.00%
10	8	6	100.00%	75.00%
11	4	1	50.00%	12.50%
12	6	7	75.00%	87.50%
13	3	7	37.50%	87.50%

Mean	5.62	5.15	70.19%	64.42%
Median	6.00	6.00	75.00%	75.00%
St Deviation	2.26	3.05	28.20%	38.14%

On average, students correctly identified 70% of arrow directions and 64% of polarities. The median score was 75% for both measures. Standard deviations were relatively high, 28% and 38%, respectively, indicating variability in performance.

Activity 2: Tax multiplier

Immediately after Activity 1, students completed the worksheet for Activity 2 (see Appendix B), which focused on the tax multiplier. Table 2 tabulates the results. Several students marked all arrow directions and polarities correctly. Overall, performance improved compared to Activity 1: 89% of arrow directions and 75% of polarities were correct on average.

Table 2: Results of completing the causal diagram for the tax multiplier.

	Correct answers		Percentage correct	
Student	Direction	Polarity	Direction	Polarity
1	8	7	100.00%	87.50%
2	8	8	100.00%	100.00%
3	4	0	50.00%	0.00%
4	8	7	100.00%	87.50%
5	5	7	62.50%	87.50%
6	8	8	100.00%	100.00%
7	8	7	100.00%	87.50%
8	8	5	100.00%	62.50%
9	8	8	100.00%	100.00%
10	5	0	62.50%	0.00%
11	7	5	87.50%	62.50%
12	8	8	100.00%	100.00%
13	8	8	100.00%	100.00%
Mean	7.15	6.00	89.42%	75.00%
Median	8.00	7.00	100.00%	87.50%
St Deviation	1.46	2.86	18.29%	35.72%

Comparison of activities

Table 3 summarizes class statistics for both activities, including the coefficient of variation (CV), calculated as SD/Mean.

Table 3. Measures of dispersion of class results.

	Direction			Polarity		
	Mean	SD	CV	Mean	SD	CV
Activity 1	70.19%	28.20%	0.40	64.42%	38.14%	0.59
Activity 2	89.42%	18.29%	0.21	75.00%	35.72%	0.48

The results indicate improvement in students' ability to identify causal directions and polarities from Activity 1 to Activity 2. Variability decreased for arrow directions (CV dropped from 0.40 to 0.21) and polarity (CV dropped from 0.59 to 0.48), suggesting that repeated exposure and instructor feedback helped students perform more consistently.

The improvement in diagram accuracy from Activity 1 to Activity 2 may reflect deeper conceptual understanding of the multipliers, but because the measures are closely tied to the format of the activity itself, it may equally reflect task-specific proficiency with the diagram exercise rather than a more general gain in understanding. Distinguishing between these possibilities would require an independent assessment of the underlying material, which this pilot did not include (see Limitations and Future Research).

Since the larger sample was unavailable at the time of this pilot, the qualitative evidence of students' satisfaction with the activities during this class, gathered through the feedback form (Form 3 in Appendix C), provides a more robust basis for demonstrating the promise of this innovative instructional approach.

Student feedback

After completing both activities, students provided written feedback, which is organized into six themes in Table 4. The full survey instrument is included in Appendix C.

Table 4. Student feedback

Themes and Interpretation	Citations of Students' Comments
Shared underlying mechanism. Students recognized that both multipliers operate through the same basic channel, changes in planned expenditure and disposable income, and in several cases described this using very similar language across independent responses.	"Both government purchases multiplier and tax multiplier affect national income through changes in planned expenditure. They both influence disposable income." "Yes, both increase consumption, national income, PE, and disposable income. They both increase economic growth." "Yes, both the government purchases multiplier and the tax multiplier are based on the same core economic principles, where the initial change in spending leads to instability in the economy."

Direct versus indirect effects. Students distinguished the government-purchases multiplier's direct effect on planned expenditure and GDP from the tax multiplier's indirect effect, which operates through its influence on consumption.	"Government purchases multiplier typically has a larger impact compared to tax multiplier due to direct increase in planned expenditure." "Government purchases multiplier directly impacts GDP at a rapid way, and not all tax multiplier goes into the GDP, as some is saved which has weaker effect." "Government purchases multiplier has a more direct impact on aggregate demand, while the tax multiplier works indirectly by increasing disposable income."
Diagrams as an aid to visualizing feedback loops. Students described the causal diagrams as helpful for seeing how a change in one variable propagates through the system.	"Yes, I like the visual diagram to see the flow of impact." "They visually explained the flow of causal relationships and feedback loops and make it easier to see how a change in one variable propagates through the economy."
Preference for dynamic over static representations. One student noted that the animated version of the diagram, shown during instructor-led explanation, was clearer than the static pencil-and-paper version completed during the activity. This theme is based on a single response and should be read as suggestive rather than representative.	"The slides displayed the animation better than pencil/paper."
Preference for formulas once the relationship was understood. One student reported that the diagrams became less useful once the underlying relationship had been grasped. Though based on a single response, this comment is notable because it parallels the expertise reversal effect described by Kalyuga et al. (2003), discussed further below.	"I think the formulas were more helpful, causal diagrams stopped helping after I understood the relationship between all of the variables."
Perceived transferability beyond the classroom. Students reflected on applying the causal-diagram approach to other domains, including work, general systems, and everyday reasoning about causation.	"Yes, they help trace "train of thoughts" while focusing on causation." "Flowcharts are very helpful to describe many systems." "I think it is possible to use them in other areas to diagram impact and see how one change impacts the cycle."

Taken together, the first two themes suggest that many students came away with a working understanding of how the two multipliers relate to one another, both in shared mechanism and in point of divergence. Rather than treating them as unrelated formulas, students appear to have organized them into a connected structure built around planned expenditure and disposable income, consistent with schema development through deliberate practice (Goffe and Wolla, 2024).

The third and fourth themes point to the diagrams' role in making the feedback loop visible, and to a preference among some students for the dynamic, animated version over the static worksheet, suggesting that sequential representations may better support schema construction than a single static image.

The fifth theme, a preference for formulas once the relationship was understood, is consistent with the expertise reversal effect described by Kalyuga et al. (2003): scaffolding that helps a novice build a new schema can become redundant once that schema is established and processed as a single automatic unit. A student reporting that the diagram "stopped helping" after the relationship was understood may be describing exactly this transition.

Finally, the sixth theme suggests some potential for transfer, though as a self-reported observation from a small number of students it should be read as a preliminary signal rather than evidence of transfer itself. These findings reinforce prior research on the benefits of causal representations for deep learning (Perkins and Grotzer, 2005; Jägerskog et al., 2019), while pointing to opportunities for refinement such as incorporating animation and clarifying when diagram-based scaffolding is most useful.

CONCLUSION

This study examined the use of causal diagrams and structural debriefing to support literacy-targeted instruction in macroeconomics. Instructors who adopt a literacy-targeted approach prioritize deeper exploration of fewer concepts, aiming to help students build coherent schemas rather than memorize isolated facts. Causal diagrams provide a visual representation of complex economic relationships, enabling students to process information by constructing and refining mental models. When integrated with conventional textbook explanations, these diagrams offer complementary perspectives that can enhance understanding economic theories. Structural debriefing serves as the pedagogical framework for introducing these diagrams into traditional classroom settings.

The primary objective of this research was to address the question: How can instructors utilize causal diagrams and structural debriefing activities for literacy-targeted instruction in a macroeconomics classroom? To explore this, we designed and implemented two activities focused on the government-purchases multiplier and the tax multiplier. Both activities incorporated causal skeletons and relevant textbook passages to guide students in completing diagrams and interpreting feedback loops.

Findings from classroom implementation indicate that reading, interpreting, and constructing causal diagrams were both engaging and challenging for students, a result consistent with prior research on structural debriefing (Capelo et al., 2024). Students demonstrated improvement in identifying causal directions and polarities across successive activities, though this improvement may reflect practice effects and increasing familiarity with the diagram format in addition to, or instead of, deeper conceptual understanding. Student feedback also revealed that animated diagrams presented during instructor-led explanations were perceived as more helpful than static diagrams, pointing to the potential benefits of dynamic visualizations.

These findings speak directly to the paper's grounding in schema theory and causal reasoning. The consistent improvement in students' ability to correctly assign causal direction and polarity across Activities 1 and 2 is consistent with the construction and refinement of an economic schema, a process in which learners progressively organize disconnected variables into a coherent causal structure through repeated,

guided practice (Ambrose et al., 2010). This aligns with research on causal reasoning showing that students often struggle to recognize complex causal relationships in new material (Perkins and Grotzer, 2005; Grotzer, 2012), and that structured practice with explicit causal representations can help address this difficulty (Jägerskog et al., 2019). Student feedback further suggests that once this schema was established, some students relied less on the diagram itself and returned to formulaic reasoning, a pattern consistent with the expertise reversal effect (Kalyuga et al., 2003).

Limitations and future research

The authors acknowledge several limitations of the study:

1) The design. The small sample size, the absence of a comparison group, and the peculiarities of the class composition make it difficult to isolate the specific effects of causal diagrams on student knowledge acquisition from the practice and instructor's guidance effects.
2) The absence of independent assessment. The paper does not present any evidence of the assessment of learning through an exam or a quiz on the topic of multipliers following in due course of the semester. Therefore, the analysis focuses primarily on short-term learning outcomes (feedback form immediately after the activity), without the evidence of long-term retention or transfer of learning.
3) The qualitative themes reported in Table 4 are drawn from the same small sample, and several are supported by only one or two student responses. These themes should therefore be read as suggestive rather than representative of a broader student population.

The findings of the impact of this instructional approach should be considered exploratory in nature due to the above limitations. More iterations of this approach and adding a control group to the design are necessary to establish its broader effectiveness.

Future research should also focus on refining instructional materials, incorporating interactive or animated representations, and exploring strategies to reduce cognitive load when students construct diagrams. The current activities model tax policy as a lump-sum tax, consistent with Mankiw's basic Keynesian Cross derivation. Future work could extend the causal diagram to a proportional income tax ($T = tY$), which would require modifying both the diagram and the resulting multiplier formula. Additionally, studies could investigate the transferability of these skills to other economic topics and real-world problem-solving contexts. Expanding this approach to larger and more diverse student populations would further validate its effectiveness and inform best practices for integrating causal reasoning into economics education.

While structural debriefing appears to be an effective complement to conventional macroeconomics instruction by making causal structures explicit, the main contribution of this paper is in the presentation and discussion of classroom activities supporting schema development and deepening conceptual

understanding, which are key goals of literacy-targeted teaching. The activities handouts are provided in the Appendices, making the paper valuable to practitioners looking for actionable teaching tools.

ENDNOTES

1. The authors gratefully acknowledge the valuable comments and suggestions received from discussants and participants at the following conferences: the 22nd St. Louis Fed Professors Conference (November 7–8, 2024); the 51st Annual Conference of the Eastern Economic Association, New York, USA (February 21, 2025); the USC Department of Economics PhD Alumni Conference, Los Angeles, CA (April 18–19, 2025); and the International System Dynamics Conference, Boston, USA (August 3–7, 2025). We also thank Ann-Sofie Jägerskog and Khalid Saeed for their insightful feedback.

2. The authors acknowledge the use of artificial intelligence tools for editing and language refinement during the preparation of this manuscript. All substantive content, analysis, and interpretations remain the responsibility of the authors.

REFERENCES

Abel, A.B., B. Bernanke, D. Croushore. (2024). Macroeconomics. 11th edition, Pearson.

Ambrose, S. A., M. W. Bridges, M. DiPietro, M.C. Lovett and M.K. Norman (2010). How learning works: Seven research-based principles for smart teaching. Hoboken, NJ: John Wiley and Sons.

Blanchard, O., A. Amighini, and F. Giavazzi. (2026). Macroeconomics. 9th Edition. Pearson.

Capelo, C. and A. Silva (2020). "Optimising the Learning Potential of Simulations Through Structural Transparency and Exploratory Guidance." Simulation & Gaming: 104687812091620.

Capelo, C., R. Pereira and J. F. Dias (2021). "Teaching the dynamics of the growth of a business venture through transparent simulations." The International Journal of Management Education 19 (3): 100549. https://www.sciencedirect.com/science/article/pii/S1472811721000987.

Capelo, C., R. Pereira and J. F. Dias (2024). "Expanding model transparency and learning potential through structural and behavioural debriefings." Systems Research and Behavioral Science. https://doi.org/10.1002/sres.3045.

Cavana RY, Dangerfield BC, Pavlov OV, Radzicki MJ and Wheat ID, Eds. 2021. Feedback Economics: Economic Modeling with System Dynamics. Springer: New York.

Collins, A., Joseph, D., and Bielaczyc, K. (2004). Design research: Theoretical and methodological issues. The Journal of the Learning Sciences, 13(1), 15–42.

Crookall, D. (2010). Serious games, debriefing, and simulation/gaming as a discipline. Simulation & Gaming: An International Journal, 41(6), 898 – 920.

Goffe, L and Wolla, S. (2024). Cognitive science teaching strategies and literacy-targeted economics complementarities. The Journal of Economic Education, 55-2, 156-165.

Grotzer, T. A. (2012). Learning Causality in a Complex World: Understandings of Consequence. Lanham, MD, R&L Education.

Jägerskog A-S. 2021a. The Affordance of Visual Tools. The Potential of Visual Representations of Pricing Facilitating an Epistemic Practice in Economics Teaching. Journal of Social Science Education 20 (1): 65-90.

Jägerskog A-S. 2021b. Using Visual Representations to Enhance Students' Understanding of Causal Relationships in Price. Scandinavian Journal of Educational Research 65 (6): 986-1003. https://doi.org/10.1080/00313831.2020.1788146.

Jägerskog A-S, Davies P and Lundholm C. 2019. Students' Understanding of Causation in Pricing: A Phenomenographic Analysis. Journal of Social Science Education 18 (3). https://www.jsse.org/index.php/jsse/article/view/1421.

Kalyuga, S., Chandler, P., and Sweller, J. (1998). Levels of expertise and instructional design. *Human Factors, 40*(1), 1–17. https://doi.org/10.1518/001872098779480587

Kalyuga, S., Ayres, P., Chandler, P., and Sweller, J. (2003). The Expertise Reversal Effect. Educational Psychologist, 38(1), 23–31. https://doi.org/10.1207/S15326985EP3801_4

Kim, Y. J., Pavlov, O. V. (2019). Game-based structural debriefing How can teachers design game-based curricula for systems thinking? Information and Learning Sciences. 120 (9/10), 567-588. https://doi.org/10.1108/ILS-05-2019-0039

Lederman, L. C. (1992). Debriefing: Toward a systematic assessment of theory and practice. Simulation & Gaming: An International Journal, 23(2), 145-160.

Maani KE and Cavana RY. 2007. Systems Thinking, System Dynamics: Managing Change and Complexity. Pearson Education New Zealand: North Shore, N.Z.

Mankiw, N. Gregory. 2025. Macroeconomics, 12th edition, Worth Publishers, Macmillan Learning, New York, NY.

McKenney, S. and Reeves, T.C. (2012), Conducting Educational Design Research, Routledge, London.

Pavlov, O. V., Y. J. Kim and C. Whitlock (2019). Food Fight: Teaching Systems Thinking and Ecosystems. Learning, Education, & Games, Volume 3: 100 Games to Use in the Classroom and Beyond. K. Schrier, ETC Press: 150-155. https://ssrn.com/abstract=3511479.

Pavlov, O. V., K. Saeed, and L. W. Robinson (2015). Improving Instructional Simulation with Structural Debriefing. Simulation and Gaming, 46 (3-4), 383–403.

Pearl, J. and D. Mackenzie (2018). The Book of Why: The New Science of Cause and Effect. New York, Basic Books.

Perkins, D. N. and T. A. Grotzer (2005). "Dimensions of Causal Understanding: the Role of Complex Causal Models in Students' Understanding of Science." Studies in Science Education 41: 117-166.

Qudrat-Ullah, H. (2020). "Improving Human Performance in Dynamic Tasks with Debriefing-Based Interactive Learning Environments: An Empirical Investigation." International Journal of Information Technology & Decision Making, 19(4): 1065-1089.

Sterman, J. D. (2000). Business Dynamics. McGraw-Hill: Boston, MA.

Sterman, J. D. (2010). "Does formal system dynamics training improve people's understanding of accumulation?" System Dynamics Review 26(4): 316-334. https://doi.org/10.1002/sdr.447.

Sweller, J. (1999). Instructional Design in Technical Areas. Melbourne, Australia. ACER Press.

Sweller, J. (2025). An Integrated Human Cognitive Architecture. *Educational Psychology Review* 37, 108 (2025). https://doi.org/10.1007/s10648-025-10089-1

APPENDIX A: Activity 1 - The government-purchases multiplier

FORM 1

Government-purchases multiplier

Date: _____ Instructor _____

Team # _____ Course: _____

Team Members:

_____

_____

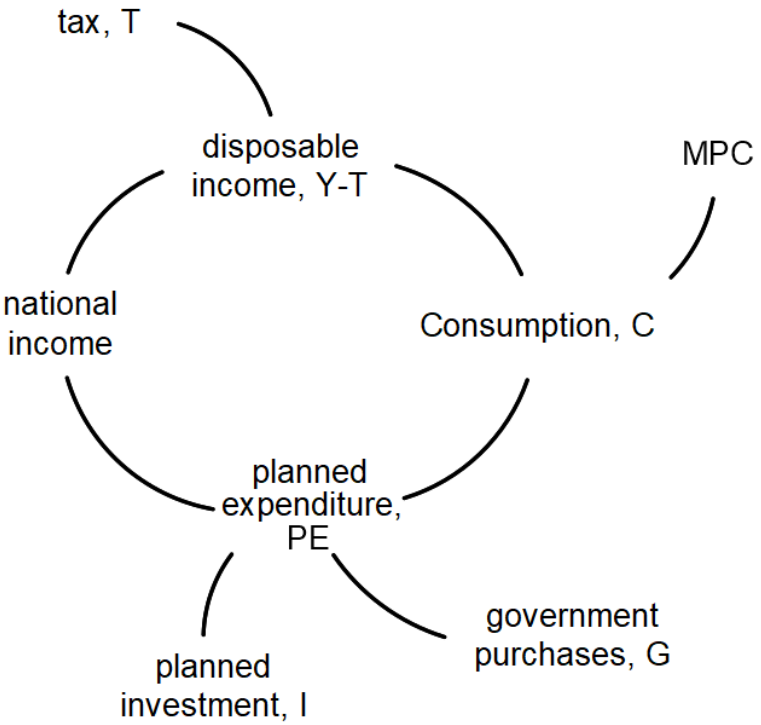
_____

Please read the following quotes from Mankiw' Macroeconomics that explain the nature of the government-purchases multiplier (pp. 314-317):

> *Assuming that the economy is closed, so that net exports are zero, we write planned expenditure PE as the sum of consumption C, planned investment I, and government purchases G:*
>
> $$PE = C + I + G$$
>
> ***An Increase in Government Purchases in the Keynesian Cross**: An increase in government purchases of raises planned expenditure by that amount …. (Source: caption for Figure 11-5, p. 317).*
>
> *…according to the consumption function $C = C(Y - T)$, higher income causes higher consumption. When an increase in government purchases raises income, it also raises consumption, which further raises income, which further raises consumption, and so on. Therefore, in this model, an increase in government purchases causes a greater increase in income.*
>
> *How big is the multiplier? To answer this question, we trace through each step of the change in income. The process begins when expenditure rises by ΔG, which implies that income rises by ΔG as well. This increase in income in turn raises consumption by $MPC \times \Delta G$, where MPC is the marginal propensity to consume. This increase in consumption raises expenditure and income once again. This second increase in income of $MPC \times G$ again raises consumption, this time by $MPC \times (MPC \times \Delta G)$, which again raises expenditure and income, and so on. This feedback from consumption to income to consumption continues indefinitely.*

Instructions: Using the provided text, complete the 'causal skeleton' graph that explains the mechanism of the government-purchases multiplier. Show causal directions and causal polarities of the

edges. Positive causal relationships between variables are drawn as positive arrows. Negative links show inverse causal relationships. Please also identify the polarity of the feedback loop. Is it a positive (reinforcing) or negative (balancing) feedback loop?



APPENDIX B: Activity 2 – The tax multiplier

FORM 2

Tax multiplier

Date: _____ Instructor _____

Team # _____ Course: _____

Team Members:

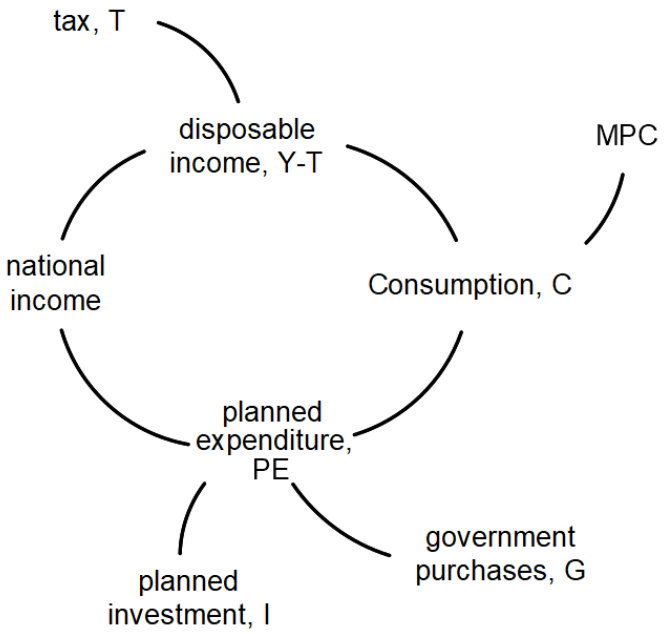
_____ _____ _____

Please read the following quote from Mankiw's Macroeconomics that explains the nature of the tax multiplier (pp. 318-319):

> *A decrease in taxes of ΔT immediately raises disposable income $Y - T$ by ΔT and, therefore, increases consumption by $MPC \times \Delta T$. For any given level of income Y, planned expenditure is now higher… Just as an increase in government purchases has a multiplied effect on income, so does a decrease in taxes. As before, the initial change in expenditure, now $MPC \times \Delta T$, is multiplied by $1/(1 - MPC)$.*

Instructions:

1. Using the above text, complete the graph below that explains the mechanism of the tax multiplier. Show causal directions and causal polarities of the edges. Positive causal relationships between variables are drawn as positive arrows. Negative links show inverse causal relationships. Please also identify the polarity of the feedback loop. Is it a positive (reinforcing) or negative (balancing) feedback loop?
2. Use the causal diagram to explain the tax multiplier effect of a decrease in taxes, ΔT.



APPENDIX C: Questionnaire

FORM 3

G and T multipliers

Date: _____ Instructor _____

Name: _____ Course: _____

Now that you've reviewed the government purchases multiplier and the tax multiplier, please answer the following questions.

1. Are there any **similarities** between the economic structures that are responsible for the government purchases multiplier and the tax multiplier?

2. Are there any **differences** between the economic structures that are responsible for the government purchases multiplier and the tax multiplier?

3. Were the causal diagrams helpful in understanding the multiplier effects? Please elaborate.

4. Did you find any part of the activity with causal diagrams confusing?

5. Do you think you can apply the causal diagrams approach to explain situations that you face at work, at home, read about in the news, etc.? Can you provide an example?

REFEREE ACKNOWLEDGMENT

The New York State Economics Association is grateful for the time spent by the following referees in recent years. Without your efforts, we would not be able to produce this journal.

1. Faisal Abbas
2. Carlos Asarta
3. Bebonchu Atems
4. James T. Bang
5. Supravat Bagli
6. Prasun Bhattacharjee
7. Gagari Chakrabarti
8. Jinyan Chang
9. Joseph Cheng
10. Brian Chezum
11. Christian Conrad
12. Jeffrey Cross
13. Michael Davis
14. Anupam Das
15. Joe Eisenhouer
16. Kousik Ganguly
17. Prithviraj Guha
18. Joanne Guo
19. Rupayan Gupta
20. L. Chukwudi Ikwueze
21. Dan Jaqua
22. Michael Jerison
23. Gisella Kagy
24. Joshua Mask
25. Michael McAvoy
26. James Murtagh
27. Michael O'Hara
28. Oleg Pavlov
29. Kasia Platt
30. Mitali Pradhan
31. Ashley Provencher
32. Jungjun Park
33. Chitrakalpa Sen
34. Julie Smith
35. Elena Smirnova
36. Jonathan Stone
37. Xuguang Sheng
38. Elias Shukralla
39. Della Lee Sue
40. Masashi Tsutsumi
41. Richard Vogel
42. Hang Yi
43. Ling Yan
44. Jenny Zhao
45. Xu Zhang