Sources of Bias in the USDA International Baseline Projections

Rabail Chandio (chandio.1@osu.edu)
Ani L. Katchova (katchova.1@osu.edu)

Please find the latest version at this link.

Working Paper
October 3, 2022

Dept. of Agricultural, Environmental, and Development Economics
The Ohio State University.

Copyright 2022 by Rabail Chandio and Ani Katchova. All rights reserved. Readers may make verbatim copies of this document for non-commercial purposes by any means, provided that this copyright notice appears on all such copies.
Abstract

USDA’s annual Agricultural Baseline Projections contribute significantly to agricultural policy in the United States, and hence their accuracy is vital. The baseline projections present a neutral policy scenario, which assume a specific macroeconomic situation and allow for the analyses of alternative policies and their micro and macroeconomic impacts in the United States. We investigate the trends and heterogeneity in the incidence of bias in the USDA International Baseline Projection reports from 2002 to 2021. The evaluation of bias as it varies geographically, temporally, and by crop and variable allows us to make inferential judgments about the sources of the projections’ bias. We begin by using the dynamic time warping algorithm to examine whether experts tend to group together the projections for certain crops across different countries, producing similar projection trends, called herding. We find that projection series for all countries in the sample are correlated with the projections for the United States. We also compute the bias in projections and decompose it by projection horizon, finding that the bias is almost the lowest for the United States while being substantially higher for other countries. Then, we assess whether the bias is higher across crops or across countries with more substantial evidence for herding behavior. We find that for Corn yield, and soybeans area harvested, herding in projection trends with the United States is associated with lower bias while for most other crops, herding is associated with significantly higher projection errors. This suggests that the baseline projection trends for other countries being similar to the United States may be bias inducing in most cases. The methods and findings of our study can help not only the USDA, but also other forecasting agencies, in improving their projection accuracy despite the lack of information.

JEL codes: C53, E17, Q11, Q13, Q14, Q18

Keywords: agricultural baselines, USDA baselines, commodity projections, time series similarity, forecast evaluation.
1 Introduction

From long-term agricultural policies that are incorporated into the law through quinquennial Farm Bills and the annual Presidential budgeting, USDA’s annual Agricultural Baseline Projections have substantial implications for various stakeholders. The baseline projections present a neutral policy scenario, which assume a specific macroeconomic situation and allow for the analyses of alternative policies and their micro and macroeconomic impacts in the United States. They help evaluate local and foreign policy scenario changes and their subsequent implications for United States’ farmers (Skorbiansky, Childs, and Hansen 2018; Langholtz et al. 2012). Therefore, any policy evaluations utilizing the baselines projections will be as useful and informative as the projections are accurate. Recently, academic research has started evaluating the accuracy of these projections, and has made initial discoveries about the incidence of bias and limited informativeness of the various USDA projections and forecasts (Bora, Katchova, and Kuethe 2022, 2021; Regmi et al. 2021; Isengildina-Massa et al. 2021; Kuethe, Hubbs, and Sanders 2018). However, from a policy perspective, understanding the source of bias is essential for minimizing it and improving the projections, which has not received much attention in the literature.

Our study addresses this gap in the literature and investigates the level of herding and its contribution towards bias in the USDA international baseline projections, where herding is defined as significant similarity in two projected series. To answer this question, we have a two-step empirical strategy. First, we examine whether experts tend to group together the projections for certain crops across different countries, defined as herding behavior, producing similar projection trends. Since projections are produced for a decade in any given year, and may be correlated in both leads and lags, we employ time-series techniques novel to our field to accommodate this time-varying correlation and quantify herding. Utilizing rich time-series data where each time period nests an entire projection series (nested time series), we use a Dynamic Time Warping (DTW) algorithm to assess the degree of herding in the baselines projections. The DTW algorithm has been adapted to various fields in recent applications (Berndt and Clifford 1994; Müller 2007; Jeong, Jeong, and Omitaomi 2011; Varatharajan et al. 2018) and is a robust way to measure similarity in two time series.

Second, we compute the bias in projections, where bias is defined as the difference
between projected values and realized values, and assess how the bias varies across crops and across countries with differing level of herding behavior. Quite reasonably, some similarities in projections may be stemming from the fact that the true realized data for those regions follow a similar trajectory. In that case, herding in the projections may occur from an informed choice, and it’s relation to bias should be evaluated with that in mind. Since we only observe the final values and cannot see any adjustments made to the initially proposed projected values, we account for this phenomenon by defining the rationality of the projections on an ex post facto basis. Simply put, the level of similarity in the realized historical data is measured also using the dynamic time warping algorithm, and, if it lies in the confidence interval of estimated similarity in the projections, we classify this as a “rational” level of herding. This allows us to observe the heterogeneity in the distribution of bias and estimate if and how irrational herding (not rational according to the defined criteria) contributes to the overall bias in the projections.

Herding is a behavioral phenomenon often observed in financial markets, when investors and experts with private information align their choices and decisions with others as a risk management strategy. It can be rational if the individuals make the choice to align their decisions with others based on superior private information, or it can be irrational if individuals ignore their private information in order to adopt similarity with others (Devenow and Welch, 1996). Behavioral finance research suggests that propensity to herd is a response to a private cost minimization strategy by individuals. As long as there is no significant cost to a single agent of agreeing with the group opinion, the majority of people choose to follow the group consensus, regardless of their individual prior beliefs (Huang et al., 2017). Moreover, when an institution (such as USDA) or a specific forecasting team is considered a single entity, the forecast behavior of experts within the institution and/or a team is affected by and aligned with the overall beliefs of the institution/team, and hence their forecasts herd together (Benchimol et al., 2020; Van Campenhout and Verhestraeten, 2010). Whether herding is rational or irrational, it increases volatility in stock and commodity markets and is suboptimal for the market (Blasco, Corredor, and Ferreruela, 2012; Wang and Wang, 2018).

Moreover, to understand how herding may occur in the baseline projections, we first clarify the process of preparing the baseline projections. Released each year by the USDA Interagency Agricultural Projections Committee, baseline projections combine
model-based values and judgment-based adjustments to these values (USDA Agricultural Projections to 2030). Experts from various committees in USDA, including the Economic Research Service, World Agricultural Outlook Board, and the Office of Chief Economist, evaluate the region specific projected values and adjust them until a point of consensus is reached among the committee. Therefore, the projections process involves a first stage where region specific projections are prepared by individual teams, and a second stage where all the regional projections are considered in unison for a global model, and the region specific values may be adjusted. We, however, only observe the finalized projection values after both stages are completed and cannot observe any adjustment. Conceptually, for the baseline projections, similarity in projections can be introduced at either of the two stages.

We find that the baseline projections for all countries are statistically significantly aligned with the projections of the United States in their trends for all crops and variables. Moreover, for most crops and variables including corn, soybeans, and wheat total consumption, this correlation is associated with significantly higher errors in projections. In addition, for the cases where the correlation in projections is rational, i.e. the realized values also exhibit similarity in trends, it is often associated with lower errors in projections. Our findings have implications for the USDA baseline experts as well as government agencies and users of the baselines reports.

We make three main contributions to the literature. First, our study identifies that the projections for all countries included in the USDA International Baseline Projection reports are correlated with the United States beyond what the realized values are for each country, which may inform USDA on another criterion that needs to be examined prior to releasing their projection reports. Second, we provide conclusive evidence that for other countries’ projections of most crop variables, correlation with the United States is associated with higher bias and lower accuracy of these countries’ projections. This informs the USDA baseline experts to assess the accuracy of the baseline models. If the excessively correlated projections were a result of the model values being correlated, then the model input or models themselves may need to be updated. However, if the models resulted in dissimilar or uncorrelated projections that were later smoothed by the baseline experts to look similar, this would suggest that the baseline committee need to reconsider their herding behavior. Third, we recognize the heterogeneity in the relationship between
projections’ bias and projections’ correlation. By highlighting the variables for which correlation reduces the error in projections, a higher accuracy in the projections can be achieved if the projections are improved. Overall, these insights can be incorporated by the team preparing the USDA baselines projections to minimize excessive similarity in the projections, which decreases their accuracy.

The remainder of the paper is organized as follows. Section 2 describes the USDA International Baseline Projections and the variables included in our study. Section 3 details the empirical strategy, which is followed by presentation and discussion of the results in section 4. Section 5 contains the concluding remarks.

2 Data

The baseline projections are one of various economic forecasts produced by several agencies in the US. We use the official USDA International Baseline Projections data from 2002 to 2021 which includes 10-year domestic (United States) and international (other countries) projections for several crops each year. We limit our analysis to corn, soybeans, and wheat for the variables area harvested, yield, imports, exports, ending stocks, and total consumption. Balance sheet equation dictating the relationship of the variables we study is as follows

\[
\begin{align*}
\text{Beginning Stocks} + \text{Production} + \text{Imports} &= \text{Exports} + \text{Total Consumption} + \text{Ending Stocks} \\
&= E + T + E
\end{align*}
\]

where the \( \text{Beginning Stocks}_t = \text{Ending Stocks}_{t-1} \), making it a redundant variable, and \( \text{Production} = \text{AreaHarvested} \times \text{Yield} \). We focus only on the variables that are identified independently, thus, not considering beginning stocks and production.

The available baseline data also includes the realized values for up to three years before the release date of the reports. We utilize these limited historical data in each year’s report to construct an annualized panel data set for realized values that are used for bias calculations and accuracy evaluations of the projections.

The baseline projections have a structure which is statistically referred to as nested time-series data, where each year nests the series of ten incremental horizon projections for 10 years or horizons into the future. A representative projection \( \hat{Y}_{rcvt} \) is the projection series for country \( r \) (belonging to an unbalanced panel of 34 countries observed
annually over the study period), for crop \( c \in \{ \text{corn, soybeans, wheat} \} \), variable \( v \in \{ \text{yield, area harvested, imports, exports, total consumption, ending stocks} \} \), and report year \( t \in \{ 2002, \ldots, 2022 \} \). \( \hat{Y}_{rct} \) is a series that has a length of \( H = 10 \), where \( h \) represents the different projection horizons such that \( \hat{Y}_{rct} = (\hat{Y}_{h_0}, \hat{Y}_{h_1}, \ldots, \hat{Y}_{h_9}) \).

The bias in the baseline projections is defined as the difference between the projection and the actual value. We employ two measures for assessing projections accuracy which are common in the literature: the logged error (LE), which can be interpreted in percentage terms):

\[
LE_{rceh} = \left( \log(\hat{Y}_{rceh}) - \log(Y_{rceh}) \right)
\]

and the root mean squared percentage error (RMSPE)

\[
RMSPE_{rceh} = \left( \frac{1}{T} \sum_{t} (100(\hat{Y}_{rceh} - Y_{rceh})/Y_{rceh})^2 \right)
\]

where \( Y_{rceh} \) is the actual value realized for the projection \( \hat{Y}_{rceh} \). \( \text{error}_{rceh} \) (\( LE_{rceh} \) or \( RMSPE_{rceh} \)) is the average error calculated over the report years \( t \) in the projections for country \( r \), crop \( c \), variable \( v \), and horizon \( h \).

3 Methods

There are two main components of our empirical analysis. First, we estimate the degree of similarity and correlation among various countries’ baseline projections using the dynamic time warping algorithm. We also compute the similarity in realized data over the study period to provide a reference to understand whether the similarity in projections is rational or not. The estimated similarity in projections is considered rational if the distance in realized value falls within the confidence interval of the estimated similarity.

Second, we use regression analysis to study the relationship between the degree of herding and the size of bias in the projections. To fully understand the heterogeneity in this relationship, we estimate it separately for all crops and variables, and we also vary the benchmark country. That allows us to observe whether the top producers of a commodity are used as benchmarks in practice when these projections reports are produced. If this is the case, then we should observe similarity in projections with the top producer of a
commodity.

3.1 Evaluating the Degree of Similarity

We begin our analysis by evaluating the differences in projections of specific countries for each crop, variable, report year, and projection horizon to estimate the degree of similarity or herding in the projections. We use a dynamic time warping algorithm to compute the distance between each set of projection series for each crop-variable-year-horizon combination (for instance, corn yield projections for all 10 horizons in the future that are included in the report year 2010) and evaluate whether the projections exhibit similar trends among all countries. The algorithm finds the minimum distance needed to make two time-series as similar as possible. We use the DTW algorithm to compute the distances between all country pairs for each crop-variable-year-horizon to determine the closest projection “neighbors” of the top country producers for each crop that have the smallest distance among all countries.

We suppress the indices $c,v,t$ since they remain the same for each pair of countries whose projections are being compared. To determine the distance between the projections for any two countries $\hat{Y}_{ri}$ and $\hat{Y}_{rj}$, we define the two time-dependent series $\hat{Y}_{ri}$ and $\hat{Y}_{rj}$ and compute an expansive local cost matrix ($LCM$) between them. The $LCM$ is populated by pairwise comparisons of each horizon’s projections for $\hat{Y}_{ri}$ with each horizon’s projections of $\hat{Y}_{rj}$, resulting in a square matrix of dimensions $10 \times 10$ since length($\hat{Y}_{ri}$) = length($\hat{Y}_{rj}$) = 10 for the 10 horizons. The $LCM$ matrix is defined as:

$$LCM(\hat{Y}_{ri}, \hat{Y}_{rj}) = \begin{pmatrix} d_{\hat{Y}_{ri},0,\hat{Y}_{rj},0} & d_{\hat{Y}_{ri},0,\hat{Y}_{rj},1} & \cdots & d_{\hat{Y}_{ri},0,\hat{Y}_{rj},9} \\ d_{\hat{Y}_{ri},1,\hat{Y}_{rj},0} & d_{\hat{Y}_{ri},1,\hat{Y}_{rj},1} & \cdots & d_{\hat{Y}_{ri},1,\hat{Y}_{rj},9} \\ \vdots & \vdots & \ddots & \vdots \\ d_{\hat{Y}_{ri},9,\hat{Y}_{rj},0} & d_{\hat{Y}_{ri},9,\hat{Y}_{rj},1} & \cdots & d_{\hat{Y}_{ri},9,\hat{Y}_{rj},9} \end{pmatrix}$$

where each matrix element $d_{\hat{Y}_{ri},a,\hat{Y}_{rj},b} = \sqrt{(\hat{Y}_{ri,a} - \hat{Y}_{rj,b})^2}$ denotes the Euclidean distance between $a^{th}$ and $b^{th}$ horizon projections of series $\hat{Y}_{ri}$ and, $\hat{Y}_{rj}$ respectively and $a, b \in h$.

We find the distance between the two projection series by defining $\phi(k)$ to be the path from $d_{\hat{Y}_{ri},0,\hat{Y}_{rj},0}$ to $d_{\hat{Y}_{ri},h,\hat{Y}_{rj},h}$ where $k = (1,1), \ldots, (H,H)$. For a given path $\phi$, we compute the Euclidean distance measuring similarity between the projections for two
countries $\hat{Y}_{ri}$ and $\hat{Y}_{rj}$ as

$$d_\phi(\hat{Y}_{ri}, \hat{Y}_{rj}) = \sum_k[(LCM(k) \times m_\phi(k))]$$

where $m_\phi(k)$ is the per-step weighting coefficient, allowing to only add the distance in horizon-specific projections that falls on a given path $\phi(k)$. At this step, all the possible paths from the beginning of $LCM$ at $k = (1,1)$ to the end at $k = (H,H)$ are computed.

Next, we find a path $\phi$ within the $LCM$ matrix that gives the minimum total distance between the projections in two countries $d_\phi(\hat{Y}_{ri}, \hat{Y}_{rj})$. Hence, we use the DTW algorithm to find the minimum distance between the two countries’ projections for the close-by horizons by solving the following optimization problem:

$$DTW(\hat{Y}_{ri}, \hat{Y}_{rj}) = \min_\phi (d_\phi(\hat{Y}_{ri}, \hat{Y}_{rj})) \quad (4)$$

We impose two constraints on the minimization problem to avoid meaningless loops or inefficient paths. First, we require monotonicity, which restricts the direction of the path taken within the $LCM$ to only increasing projection horizon for at least one of the projection series:

$$\phi : \phi(k + 1) \geq \phi(k) \quad (5)$$

Second, we impose the Sakoe-Chiba window constraint to restrict the path that can be taken to a window $h_{\text{window}}$ around the principal diagonal of the $LCM$. We use a window $h_{\text{window}}$ as the only valid points of the $LCM$ where a path $\phi$ can be made in the range:

$$\lfloor (h_a, h_b - h_{\text{window}}), (h_a, h_b + h_{\text{window}}) \rfloor \quad (6)$$

for all horizons $(h_a, h_b)$ along the principal $LCM$ diagonal. This constraint allows only reasonable lead and lag relationships of the horizons to be considered while two projection series are being compared. We set $h_{\text{window}} = 2$ restricting projection $\hat{Y}_{ri}$ for, say, horizon 5 to only be compared to projection $\hat{Y}_{rj}$ for another country from horizon 3 to 7 allowing the algorithm to detect at max a 2-year lead or a 2-year lag relationship between the two projection series. This approach is more general than restricting comparisons between countries’ projections to be only for the same horizon. To ensure comparability across different countries’ projections, we scale the baseline projections data using z-score normalization for each projection series being compared.
The distance calculation detailed above is repeated for each crop $c$, variable $v$, report year $t$, and the projection series of each country have the length of horizon $h$. To compute the overall distance in projections of two countries, we average the DTW distances between projections for two countries $\hat{Y}_r$, and $\hat{Y}_j$ over all report years $t$:

$$
\text{distance}(\hat{Y}_{r,cvt}, \hat{Y}_{j,cvt}) = \frac{1}{\text{length}(t)} \sum_{t=2002}^{2021} (DTW(\hat{Y}_{r,cvt}, \hat{Y}_{r,cvt}))
$$

(7)

where $\text{distance}(\hat{Y}_{r,cvh}, \hat{Y}_{j,cvh})$ gives average difference in projections of two countries for each crop $c$ and variable $v$ and $DTW(\hat{Y}_{r,cvt}, \hat{Y}_{j,cvt})$ refers to distance from the minimization problem described in equations (4), (5), and (6). We compute the standard errors for the computed distance using the bootstrapping method with 250 replications.

If a country’s distance from a benchmark country is statistically indistinguishable from zero, we would infer that its projections are correlated with those of the benchmark country, in other words, the two countries are similar in trends. Repeating this process for all crops and variable combinations and all benchmark countries provides enough information to evaluate whether and where herding behavior of countries’ projections occurs.

### 3.2 Relationship between Bias and Herding

We estimate the relationship between similarity in projection trends and the bias in the baseline projections. This allows us to answer whether herding the projections for various countries is desirable in the case of limited information. Moreover, we can evaluate the heterogeneity in the impact of herding on bias across crops, variables, horizons, and report years. Using different countries which are the top producers of these crops (United States, Brazil, or China) as benchmark countries for herding in the projections, we estimate the following equation:

$$
LE_{rh} = \beta_0 + \beta_1 \log(DistanceFromBase)_r + \beta_2 CorrelatedWithBase_r \\
+ \beta_3 DistanceIsRational_r + \beta_4 (CorrelatedWithBase_r \times DistanceIsRational_r) \\
+ \epsilon_{rh}
$$

(8)

for each crop and variable separately, using the log error (LE) calculated in equation (2). $\log(DistanceFromBase_r)$ is the log of computed DTW distance of country $r$’s
projections from the benchmark country $r_{base}$ for each projection horizon. This is the distance calculated in equation (4). $CorrelatedWithBase_r$ is an indicator variable taking a value of 1 if, on average, country $r$’s projections for the given crop and variable are correlated with the benchmark country $r_{base}$, and 0 otherwise. $DistanceIsRational_r$ is an indicator variable that takes a value of 1 if the distance in the realized series lies within the confidence interval of the average distance in the projections of country $r$ from the benchmark country. The final term is the interaction of the two indicator variables, reflecting the effect of rational correlation in projections on the error in country $r$’s projections.

The coefficient on $\log(DistanceFromBase_r)$ shows the $\beta_1$% change in the projection error associated with 1% increase in the distance of country $r$’s projections with the benchmark country $r_{base}$. If this value is significant and positive, it implies that herding with the benchmark country is associated with lower errors, while a significant negative estimate suggests that herding with the specified benchmark country may be causing higher projection errors. Moreover, the coefficient on $CorrelatedWithBase_r$ shows ($\beta_2 \times 100$)% percent change in error if the distance in the projections of country $r$ is significantly similar to the projections of benchmark country $r_{base}$. Similarly, ($\beta_3 \times 100$)% is the change in projection error if realized values is within the confidence interval of the distance in the projections, i.e., distance is rational. A significant positive estimate for $\beta_2$ and $\beta_3$ is interpreted as herding being associated with higher errors, and rational distance in projections being associated with higher errors, respectively.

We estimate two versions of equation (8). When the United States is set as the benchmark country, we only include the variables $\log(DistanceFromBase_r)$ and $DistanceIsRational_r$. As we will see in the results section, that is because all other countries’ projections are correlated with the United States, and hence the other two remaining variables are redundant in the regression. Since all countries’ projections are correlated with the United States projections, the variable $DistanceIsRational_r$ can be interpreted as the change in error by switching from an irrational correlation to a rational correlation in the regressions where United States is the benchmark region.
4 Results

We first present the results of the dynamic time warping algorithm that measures the distance in projections of various countries compared to the projections of select benchmark countries i.e., United States, Brazil, and China. These three regions are chosen as the benchmark first because they are among the leading global producers for the crops we are focusing on: corn, soybeans, and wheat. The results are categorized by crop (corn, soybeans, and wheat) and variable (yield, area harvested, ending stocks, total consumption, imports, and exports). We also depict the errors in the projections for each country, displaying them separately for projection horizon. So, the top panels of 1 through 18 correspond to similarity estimates from equation 7 and the bottom panel is a visual representation of error calculations as per equation 2. The similarity measures are estimates, so we also depict the 95% confidence interval for all of them. Therefore, if the confidence intervals for estimated distance contains 0, we conclude that the projections for the country on the horizontal axis are statistically significantly similar to the projections for the benchmark country. We have estimated the similarity for all countries in our sample, but display only twelve countries in the figures, since they are among the major producers of the three crops included in our analysis. Figures 1 to 6 correspond to corn, figures 7 to 12 represent soybeans, and figures 13 to 18 contain the results for wheat.

Our results show strong correlation in the baseline projections trends of other countries with the United States. For all the countries in our sample, all confidence intervals for the distance from the United States contain 0 regardless of the magnitude of the point estimate. That implies that projections for all crops and variables are significantly similar to the United States (their distance from the United States is indistinguishable from 0). On the other hand, there is reasonable variation in similarity of projections with the other benchmark countries (Brazil and China). For instance, in figure 1 distance in projections between Brazil and United States overlaps zero (projections are similar) while distance between Brazil and China is significantly greater than zero (projections are not similar). Ukraine and Indonesia are two other countries in figure 4 with significant similarity to the United States while being significantly distant from Brazil.

The most notable result from these set of figures is that the USDA baseline projections do not significantly differ from the United States compared to other sample countries for
any crop or variable considered in the analysis, while reasonable dissimilarity exists when Brazil or China are set as the benchmark country. However, the distances in realized values of different countries from the United States are often greater in magnitude than the estimated distance in projections. For select variables, the distance in realized values are outside the 95% confidence interval of estimated distance in projections (i.e., distances are irrational), the exception to this occurs for variables with very large confidence intervals (see corn area harvested—figure 2 and corn total consumption—figure 3).

Moreover, projection error trends in the bottom panels of figures 1 through 18 show the error in projections for the twelve countries depicted in the top panels. This is the average error over the study period for each projection horizon, marked on the horizontal axis. A higher projection horizon means that projections are being made further into the future. We observe that while the projection errors for the United States remain among the lowest, the errors for most other regions increase drastically as the projection horizon increases. Relatively, yields are projected with the most certainty since there are no significant changes in a country’s productivity from year to year, so the errors for yields are relatively smaller than other variables for all three crops across all countries. For the United States, errors in corn, soybeans, and wheat yields remain lower than 10% in magnitude for all horizons indicative of accurate yield projections. Contrarily, yield projections errors for other regions (see Ukraine and Brazil, for instance) are lower than errors in their projections for other variables, but are visibly much higher than the errors for the United States.

While the overall errors for other variables are higher in magnitude compared to errors in yields, the United States consistently continues to have lower errors relative to other countries in almost all the figures. It is reasonable to wonder, then, if the projections for other regions being herded with the United States has any contribution to this phenomenon. USDA can project the values for the United States seemingly very well, but there is still room for improvement. Therefore, a higher level of accuracy in the projections of other countries will make for a better understanding of the global environment and can lead to an improvement in domestic projections. To see why it might be so, consider the balance sheet relationship depicted in equation (1). The crop specific balance sheet equation holds for each country individually in the projections, and it also holds in summation for the whole world. Therefore, we now look at the role of
herding with the United States, and other commodity leaders (Brazil and China) in the
determination of errors.

4.1 Relationship between Herding and Bias

Results for estimation equation (8) for each of the three base countries—United States,
China, and Brazil, respectively—are presented separately in tables 1 to 3. Equation (8)
estimates the association between error in projections and the distance in projections of
a region from the benchmark country. Each column and crop panel in shows estimation
results corresponding to equation (8) with the benchmark country specified in the table
caption.

Table 1 shows the results with the United States as the benchmark country. The
coefficient on $\text{Distance from Base}_r$ (logged) measures the percentage change in the error
if the projections are 1% farther from the United States, which serves as the base country.
The first row implies that corn imports and corn ending stocks projections being more
distant from the United States are associated with significantly lower projection errors
among other countries (i.e., a negative significant coefficient on $\text{Distance from Base}_r$
(logged)). That is, decreasing the correlation with the United States in these projections
for all countries, which on average are all significantly correlated with the United States,
is associated with decreases in their errors. We observe a similar relationship for soybeans
imports, and ending stocks, and all wheat projections except imports. Herding, in these
cases, is related to higher errors.

Moreover, the coefficient on $\text{Distance is Rational}_r$ shows the $\beta \times 100$ percent change
in error if the distance in the projected values is rational. Recall that all the projections
are correlated with the United States so, in this specific case, this coefficient can be
interpreted as the change in error by switching from an irrational correlation to a rational
correlation. Therefore, rational correlation in the projections of corn total consumption
is associated with a 10.8% lower projection error. Except for corn total consumption,
all other corn variables have higher projection errors even when the correlation with the
United States in projections is rational. That is, herding, even if it is rational, does not
seem to be associated with lower errors.

For soybeans, projection errors in imports, and ending stocks decrease as the projec-
tions for each country differ more from the United States. Soybeans panel, coefficients
on *Distance from Base* \(_r\) \((\text{logged})\). Moreover, all variables for soybeans have a lower error associated with rational correlation from the United States in projections, except for exports (see coefficient on *Distance is Rational* \(_r\)). *Distance is Rational* \(_r\) is related with a higher error in soybeans exports. That means that higher error is observed in the countries where the realized values are as correlated as the projections. This is because for soybeans ending stocks, there is less variation in the variable *Distance Is Rational* \(_r\) (confidence intervals for estimated distance are very high so this variable offers limited information), while the overall errors are higher for almost all countries - see figure 12. Therefore, only a few outlying countries, for which the distance is not rational, are identifying this coefficient, and they happen to have lower errors.

For wheat, we observe that higher distance from the United States, i.e., less herding in projections, is significantly associated with lower projection errors for all variables except for imports (Wheat panel, coefficients on *Distance from Base* \(_r\) \((\text{logged})\)). The *Distance is Rational* \(_r\) variable, on the other hand, has a significant negative coefficient for wheat total consumption but a positive coefficient for wheat yield and exports. This implies that if the distance in projections from the United States is close to the distance in realized values for wheat yields and exports, it is still related to higher error in the projections. Recall that all projections are correlated with the United States for all variables. This variable, then, depicts the effect of rational versus irrational correlation with the United States. Since higher distance from the United States reduced error in wheat projections regardless of rationality, the positive coefficient on *Distance is Rational* \(_r\) for wheat yield and exports may not be as informative.

Overall, there are two main takeaways from these results. First, on average, countries whose projections are more distant from the United States are associated with lower projection errors for most variables across all three crops, even if marginally so. This suggests that the projections of other countries, that are following similar trends to the United States projections, is associated with higher bias in these countries’ projections in most cases. The exceptions are corn yield, and soybean area harvested, where larger distance from the United States in projections is significantly associated with higher projection errors for these countries’ variables. Second, in when the distance in projections for a country from the United States is rational, in addition to being correlated, whether the error is higher or lower in the projections depends on the distribution of the specific
variable.

We also use Brazil and China as the base countries because they are major crop producers and trade partners of the United States. Therefore, we repeat the analysis to also assess the correlation in projections by setting China and Brazil as benchmark countries instead of the United States. Tables 2 and 3 show the estimation results from equation (8) by setting China and Brazil as base countries, respectively. Since the projections for all countries are statistically correlated with the United States in their trends, using other top producers as the base country allows us to evaluate whether the other major producer countries offer an alternate error minimizing approach to herding. While limited information of the baseline team may result in projections for other countries to follow and correlate with the projections of the major producers, it does not necessarily mean that the United States trends are applied globally. It is also reasonable to expect that for some crops and variables, following the projection trends of China or Brazil in projections may reduce errors because of the massive global contribution of these countries, while for others it may not. Our results depict that heterogeneity.

For all crops and variables in table 2, increasing a country’s projection distance from China reduces the projection errors on average. Four crop-variable combinations are an exception to this, namely corn yield and area harvested, soybeans exports and wheat imports, which have a positive significant coefficient for Distance from Base, (logged). Moreover, a definitive correlation in projection trends with China is associated with significantly lower errors in corn ending stocks, and wheat yield and area harvested. This suggests that, on average, projections trends for some countries with probably less information follow China in projections for corn ending stocks, and wheat yield and area harvested, and this alignment seems to reduce their error.

On the other hand, the most interesting result to note in table 3 where Brazil is the benchmark is that only three of the coefficients for the variable “Correlated with Base,“ are negative and significant for any crops or variables. Contrarily, corn are harvested, imports and total consumption, soybean yield, area harvested and total consumption, along with wheat yield and imports all depict a significant positive relationship between other countries’ correlation with Brazil and higher projection errors.

The most important thing to note from the results with China and Brazil as benchmark regions is that the positive or negative association with errors when distance in
projections from either of these countries increases is the same. That is, other countries’ projections being closer to either China or Brazil for corn yield is associated with higher errors. Moreover, for the variables that are significant across all three tables (tables 1, 2, and 3), the signs for the coefficient on \( \text{Distance from Base, \ (logged)} \) is the same. This is reassuring for us because, as established before, projections for China and Brazil are both significantly similar to the United States. But, using China and Brazil as base regions separately in our estimation allows us to observe exactly what level of similarity is helpful in reducing the errors while when does it lower the accuracy of our projections.

For instance, consider corn area harvested. The coefficient on \( \text{Distance from Base, \ (logged)} \) is insignificant when United States is the benchmark country, but it is positive and significant when China or Brazil are the benchmark countries. That suggests that, on average, the direction, and size of other countries’ corn area harvested relates less to the United States, more to China, and more still to Brazil. Since Brazil is the largest corn producer globally, perhaps the level of production (coming from yield and area harvested together) that is projected in Brazil is based on more definitive information about global demand. Therefore, if Brazil is projected to increase its corn production over the projection horizon, probably the other countries may be doing the same for reasons that cannot be observed by the team preparing the projection for those regions. As such, making the corn yield and area harvested projections closer to Brazil rather than the United States may reduce error when there is incomplete information. Similarly, the negative significant estimates for the coefficient on \( \text{Distance from Base, \ (logged)} \) for wheat yield and area harvested are indicative that none of the three countries selected as base regions here should be chosen as the global trend leader for wheat production projections. Despite being among the top producers, China and United States should not lead the global projection trends. Historically, India and Russia have been as important, if not more, than the United States for wheat production. Therefore, it is possible that, for wheat, those countries’ trends are more representative of the global market.

5 Conclusion

While the USDA International Baseline Projections are prepared through a combination of model-based values and expert analysts’ judgements, the baseline projections for all
other countries are found to have an underlying correlation with the United States. Our results show that this is associated with lower bias for a few crop variables, but contributes to higher error in other cases, compromising the overall accuracy of the projections. Given the importance of the baseline projections in domestic agricultural policy, it is imperative to identify where the projection correlation among countries beyond the correlation in the realized values is increasing bias and reducing accuracy.

We employ various methods to identify the correlation in the projections of different countries, assess their degree of accuracy, and map the relationship between projections’ similarity in terms of correlation and projections’ error. Our results show that select variables that are grouped together in their projection trends are associated with reduced errors. Corn yield, and soybeans area harvested are the variables where similarity in projection trends with the United States is associated with more accurate projections for the other countries. Among other crop variables, our results show that correlation in projection trends is significantly associated with lower accuracy of the projections in corn imports and ending stocks, soybeans imports and ending stocks, and all wheat variables except imports.

These findings can be used by the team preparing the USDA baselines projections by checking that the projections’ correlations does not exceed realized values’ correlations, as this may decrease their accuracy. The heterogeneity observed in our results provides a reasonable starting point for the team preparing the baseline projections to get closer to the source of bias in the projections. Utilizing data on the projections made from models compared to the projections after adjustments can offer valuable insights on where and why herding occurs, and in which cases does it reduce bias.
References


Figure 1: Corn Yield—Correlation Estimates and Error Calculations
Figure 2: Corn Area Harvested—Correlation and Error Calculations
Figure 3: Corn Total Consumption—Correlation and Error Calculations
Figure 4: Corn Ending Stocks—Correlation and Error Calculations
Figure 5: Corn Imports—Correlation and Error Calculations
Figure 6: Corn Exports—Correlation and Error Calculations
Figure 7: Soybeans Yield—Correlation and Error Calculations
Figure 8: Soybeans Area Harvested—Correlation and Error Calculations
Figure 9: Soybeans Total Consumption—Correlation and Error Calculations
Figure 10: Soybeans Ending Stocks—Correlation and Error Calculations
Figure 11: Soybeans Imports—Correlation and Error Calculations
Figure 12: Soybeans Exports—Correlation and Error Calculations
Figure 13: Wheat Yield—Correlation Estimates and Error Calculations
Figure 14: Wheat Area Harvested—Correlation and Error Calculations
Figure 15: Wheat Total Consumption—Correlation and Error Calculations
Figure 16: Wheat Ending Stocks—Correlation and Error Calculations
Figure 17: Wheat Imports—Correlation and Error Calculations
Figure 18: Wheat Exports—Correlation and Error Calculations
### Table 1: Distance and Accuracy - Base country is United States

<table>
<thead>
<tr>
<th>Variable</th>
<th>Yield</th>
<th>Area Harvested</th>
<th>Imports</th>
<th>Exports</th>
<th>Total Consumption</th>
<th>Ending Stocks</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Corn</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Distance from Base, (logged)</td>
<td>0.0106*</td>
<td>-0.0617</td>
<td>-0.7131***</td>
<td>0.152</td>
<td>0.0093</td>
<td>-0.4544***</td>
</tr>
<tr>
<td></td>
<td>(0.0061)</td>
<td>(0.0551)</td>
<td>(0.1654)</td>
<td>(0.1344)</td>
<td>(0.0161)</td>
<td>(0.1024)</td>
</tr>
<tr>
<td>Distance is Rational,</td>
<td>0.407***</td>
<td>1.3221***</td>
<td>-0.208</td>
<td>-0.1079***</td>
<td>0.366</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0644)</td>
<td>(0.3034)</td>
<td>(0.1950)</td>
<td>(0.0322)</td>
<td>(0.2273)</td>
<td></td>
</tr>
<tr>
<td><strong>Soybeans</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Distance from Base, (logged)</td>
<td>3e-04</td>
<td>0.2593**</td>
<td>-0.7046**</td>
<td>0.0913</td>
<td>-0.2282</td>
<td>-0.1836*</td>
</tr>
<tr>
<td></td>
<td>(0.0113)</td>
<td>(0.1284)</td>
<td>(0.3282)</td>
<td>(0.1464)</td>
<td>(0.1544)</td>
<td>(0.1110)</td>
</tr>
<tr>
<td>Distance is Rational,</td>
<td>-0.0628***</td>
<td>-0.8825***</td>
<td>-0.4732</td>
<td>0.3019*</td>
<td>-0.0191</td>
<td>-1.1144*</td>
</tr>
<tr>
<td></td>
<td>(0.0238)</td>
<td>(0.1490)</td>
<td>(0.8614)</td>
<td>(0.1899)</td>
<td>(0.5812)</td>
<td></td>
</tr>
<tr>
<td><strong>Wheat</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Distance from Base, (logged)</td>
<td>-0.046***</td>
<td>-0.0734***</td>
<td>0.0676</td>
<td>-0.4631***</td>
<td>-0.0205**</td>
<td>-0.1793***</td>
</tr>
<tr>
<td></td>
<td>(0.0112)</td>
<td>(0.0263)</td>
<td>(0.0658)</td>
<td>(0.1245)</td>
<td>(0.0087)</td>
<td>(0.0341)</td>
</tr>
<tr>
<td>Distance is Rational,</td>
<td>0.0753***</td>
<td>0.0243</td>
<td>0.053</td>
<td>0.4419*</td>
<td>-0.0275*</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0216)</td>
<td>(0.0553)</td>
<td>(0.1588)</td>
<td>(0.2350)</td>
<td>(0.0150)</td>
<td></td>
</tr>
</tbody>
</table>

This table shows the estimation results for equation \( \log(\text{error})_{rh} = \beta_0 + \beta_1 \log(\text{Distance \_From \_Base})_r + \beta_3 \text{Distance \_is \_Rational}_r + \epsilon_{rh} \) for each crop and variable. Since all countries are significantly correlated with the United States in their projections, the other terms are irrelevant for estimation with the United States as the base country. Each column and panel shows the results for a separate regression for the crop-variable labeled in the table. Parentheses contain robust standard errors. *** \( p<0.01 \), ** \( p<0.05 \), * \( p<0.1 \).
Table 2: Distance and Accuracy - Base country is China

<table>
<thead>
<tr>
<th>Variable</th>
<th>Yield</th>
<th>Area Harvested</th>
<th>Imports</th>
<th>Exports</th>
<th>Total Consumption</th>
<th>Ending Stocks</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Corn</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Distance from Base, $r$ (logged)</td>
<td>0.0282**</td>
<td>0.0514***</td>
<td>-0.1048*</td>
<td>-0.4449**</td>
<td>-0.0874***</td>
<td>-0.4445***</td>
</tr>
<tr>
<td>Correlated with Base, $r$</td>
<td>-0.0234</td>
<td>0.159***</td>
<td>-0.0369</td>
<td>0.3917***</td>
<td>0.024</td>
<td>-0.1218**</td>
</tr>
<tr>
<td>Distance is Rational, $r$</td>
<td>0.1455***</td>
<td>-0.0693***</td>
<td>0.3864***</td>
<td>-1.6049***</td>
<td>0.0169</td>
<td>0.6603***</td>
</tr>
<tr>
<td>Correlated with Base, $r$, $r$</td>
<td>0.1203***</td>
<td>(0.0419)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Soybeans</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Distance from Base, $r$ (logged)</td>
<td>-0.0118</td>
<td>0.019</td>
<td>-0.2379**</td>
<td>0.7377***</td>
<td>-0.2867**</td>
<td>-0.1108</td>
</tr>
<tr>
<td>Correlated with Base, $r$</td>
<td>0.0733***</td>
<td>0.0837***</td>
<td>(0.1201)</td>
<td>(0.2701)</td>
<td>(0.1266)</td>
<td>(0.0817)</td>
</tr>
<tr>
<td>Distance is Rational, $r$</td>
<td>-0.0542*</td>
<td>0.0759</td>
<td>0.4567*</td>
<td>-1.9995***</td>
<td>0.0591**</td>
<td>-0.0434</td>
</tr>
<tr>
<td>Correlated with Base, $r$, $r$</td>
<td>6e-04</td>
<td>(0.1534)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Wheat</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Distance from Base, $r$ (logged)</td>
<td>-0.0338***</td>
<td>-0.1356***</td>
<td>0.2513***</td>
<td>-0.0261</td>
<td>-0.0398***</td>
<td>0.0072</td>
</tr>
<tr>
<td>Correlated with Base, $r$</td>
<td>-0.0898***</td>
<td>-0.0586**</td>
<td>0.3498***</td>
<td>-0.0062</td>
<td>(0.0136)</td>
<td>(0.0614)</td>
</tr>
<tr>
<td>Distance is Rational, $r$</td>
<td>-0.0455</td>
<td>0.0629***</td>
<td>0.0638*</td>
<td>-0.1599</td>
<td>0.0772***</td>
<td>-0.4189**</td>
</tr>
<tr>
<td>Correlated with Base, $r$, $r$</td>
<td>0.091***</td>
<td>(0.0247)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

This table shows the estimation results for equation $log(error_{rh}) = \beta_0 + \beta_1 log(Distance From Base)_r + \beta_2 Correlated with Base_r + \beta_3 Distance is Rational_r + \beta_4 Correlated with Base_r \times Distance is Rational_r + \epsilon_{rh}$ for each crop and variable. Each column and panel shows the results for a separate regression for the crop-variable labeled in the table. Parentheses contain robust standard errors. ***, p<0.01, **, p<0.05, *, p<0.1
Table 3: Distance and Accuracy - Base country is Brazil

<table>
<thead>
<tr>
<th>Variable</th>
<th>Yield</th>
<th>Area Harvested</th>
<th>Imports</th>
<th>Exports</th>
<th>Total Consumption</th>
<th>Ending Stocks</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Corn</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Distance from Base &lt;sub&gt;r&lt;/sub&gt; (logged)</td>
<td>0.0524**</td>
<td>0.0812***</td>
<td>-0.1041</td>
<td>-0.1947</td>
<td>-0.0775***</td>
<td>-0.3701***</td>
</tr>
<tr>
<td></td>
<td>(0.0207)</td>
<td>(0.0213)</td>
<td>(0.1129)</td>
<td>(0.2467)</td>
<td>(0.0144)</td>
<td>(0.0934)</td>
</tr>
<tr>
<td>Correlated with Base &lt;sub&gt;r&lt;/sub&gt;</td>
<td>-0.0024</td>
<td>0.1725***</td>
<td>0.5654***</td>
<td>-0.1321</td>
<td>0.097***</td>
<td>-0.63***</td>
</tr>
<tr>
<td></td>
<td>(0.0192)</td>
<td>(0.0156)</td>
<td>(0.0794)</td>
<td>(0.1840)</td>
<td>(0.0256)</td>
<td>(0.1088)</td>
</tr>
<tr>
<td>Distance is Rational &lt;sub&gt;r&lt;/sub&gt;</td>
<td>0.0488</td>
<td>-0.1958***</td>
<td>0.6379***</td>
<td>0.9218**</td>
<td>0.1325***</td>
<td>0.1298</td>
</tr>
<tr>
<td></td>
<td>(0.0379)</td>
<td>(0.0571)</td>
<td>(0.1753)</td>
<td>(0.3928)</td>
<td>(0.0313)</td>
<td>(0.2285)</td>
</tr>
<tr>
<td><strong>Soybeans</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Distance from Base &lt;sub&gt;r&lt;/sub&gt; (logged)</td>
<td>0.0059</td>
<td>0.0723**</td>
<td>-0.7632***</td>
<td>-0.075</td>
<td>-0.2403*</td>
<td>-0.0607</td>
</tr>
<tr>
<td></td>
<td>(0.0223)</td>
<td>(0.0324)</td>
<td>(0.2090)</td>
<td>(0.1065)</td>
<td>(0.1417)</td>
<td>(0.0700)</td>
</tr>
<tr>
<td>Correlated with Base &lt;sub&gt;r&lt;/sub&gt;</td>
<td>0.0748***</td>
<td>0.3488***</td>
<td>0.2090</td>
<td>0.1065</td>
<td>0.1437***</td>
<td>0.2773</td>
</tr>
<tr>
<td></td>
<td>(0.0116)</td>
<td>(0.0866)</td>
<td>(0.1723)</td>
<td>(0.0395)</td>
<td>(0.2126)</td>
<td></td>
</tr>
<tr>
<td>Distance is Rational &lt;sub&gt;r&lt;/sub&gt;</td>
<td>-0.0761**</td>
<td>-0.3142***</td>
<td>-0.5656</td>
<td>0.3293*</td>
<td>0.1437***</td>
<td>-0.4239</td>
</tr>
<tr>
<td></td>
<td>(0.0323)</td>
<td>(0.0777)</td>
<td>(0.5900)</td>
<td>(0.1850)</td>
<td>(0.0395)</td>
<td>(0.2773)</td>
</tr>
<tr>
<td>Distance is Rational &lt;sub&gt;r&lt;/sub&gt; × Correlated with Base &lt;sub&gt;rh&lt;/sub&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-0.3427</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.2126)</td>
</tr>
<tr>
<td><strong>Wheat</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Distance from Base &lt;sub&gt;r&lt;/sub&gt; (logged)</td>
<td>-0.0246**</td>
<td>-0.1197***</td>
<td>0.1517***</td>
<td>-0.3792*</td>
<td>-0.0582***</td>
<td>0.0024</td>
</tr>
<tr>
<td></td>
<td>(0.0095)</td>
<td>(0.0317)</td>
<td>(0.0572)</td>
<td>(0.2111)</td>
<td>(0.0147)</td>
<td>(0.0603)</td>
</tr>
<tr>
<td>Correlated with Base &lt;sub&gt;r&lt;/sub&gt;</td>
<td>0.0185*</td>
<td>-0.1008*</td>
<td>0.3979***</td>
<td>0.0268</td>
<td>-0.2024*</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0108)</td>
<td>(0.0554)</td>
<td>(0.0684)</td>
<td>(0.0172)</td>
<td>(0.1111)</td>
<td></td>
</tr>
<tr>
<td>Distance is Rational &lt;sub&gt;r&lt;/sub&gt;</td>
<td>0.022</td>
<td>0.053</td>
<td>0.1002*</td>
<td>0.0859</td>
<td>0.0405***</td>
<td>-0.2521*</td>
</tr>
<tr>
<td></td>
<td>(0.0183)</td>
<td>(0.0493)</td>
<td>(0.0526)</td>
<td>(0.4425)</td>
<td>(0.0154)</td>
<td>(0.1407)</td>
</tr>
<tr>
<td>Distance is Rational &lt;sub&gt;r&lt;/sub&gt; × Correlated with Base &lt;sub&gt;rh&lt;/sub&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.0065</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.0288)</td>
</tr>
</tbody>
</table>

This table shows the estimation results for equation $\log(error)_{rh} = \beta_0 + \beta_1 \log(Distance\ From\ Base) + \beta_2 Correlated\ with\ Base + \beta_3 Distance\ is\ Rational + \beta_4 (Correlated\ with\ Base \times Distance\ is\ Rational) + \epsilon_{rh}$ for each crop and variable. Each column and panel shows the results for a separate regression for the crop-variable labeled in the table. Parentheses contain robust standard errors. **p<0.01, *p<0.1