Qualitative crop condition survey reveals spatiotemporal production patterns and allows early yield prediction

Santiago Beguería and Marco P. Maneta

*Department of Soil and Water, Estación Experimental de Aula Dei–Consejo Superior de Investigaciones Científicas (EEAD-CSIC), 50014 Zaragoza, Spain; †Geosciences Department, University of Montana, Missoula, MT 59812; and ‡Department of Ecosystem and Conservation Sciences, W.A. Franke College of Forestry and Conservation, University of Montana, Missoula, MT 59812

Published: October 12, 2019

Raw data are gathered at the end of each week, and the official CPC report is issued at 4 PM Eastern Standard Time on the first business day of the following week.

The US Department of Agriculture (USDA) Crop Progress and Condition (CPC) report survey between 1987 and 2019 for four major crops across the US, and show how to transform the original qualitative data into a continuous, probabilistic variable better suited to quantitative analysis. Although the CPC reflects the subjective perception of many surveyors at different locations, the underlying models that describe the reported crop status are statistically robust and maintain similar characteristics across different crops, exhibit long-term stability, and have nation-wide validity. We discuss the origin and interpretation of existing spatial and temporal biases in the survey data. Finally, we propose a quantitative Crop Condition Index based on the CPC survey and demonstrate how this index can be used to monitor crop status and provide earlier and more precise predictions of crop yields than official USDA forecasts released midseason.

The authors declare no competing interest.

Author contributions: S.B. designed research; S.B. and M.P.M. performed research; S.B. and M.P.M. analyzed data; and S.B. and M.P.M. wrote the paper.

This open access article is distributed under Creative Commons Attribution-NonCommercial-NoDerivatives License 4.0 (CC BY-NC-ND).

Data deposition: The CCI dataset and the R code of the yield prediction model may be downloaded from the institutional repository of Consejo Superior de Investigaciones Científicas (CSIC): http://hdl.handle.net/10261/120195.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.1917774117/-/DCSupplemental.

First published July 16, 2020.

PNAS | August 4, 2020 | vol. 117 | no. 31 | 18317–18323

Significance

We show how subjective information from qualitative crop rating surveys conducted weekly by the USDA can be transformed into a continuous crop condition index that integrates meteorological, agronomic, physiological, technological, and management factors. This index allows comparison of crop conditions between years and locations and provides superior information that enhances yield forecasting models. The proposed methodology can be used to develop better agricultural drought monitoring and early warning systems that can anticipate production anomalies and inform decision making.
set the center of the scale in the middle of the good category, which is considered neutral according to the USDA instructions *(SI Appendix, Table S1)*. After a relatively normal or even positive start of season at the end of May and early June, the survey responses become increasingly negative in most states as the prevailing drought conditions and unusually hot weather negatively affected corn crops.

Due to their capacity to inform about ongoing or developing crop anomalies that may affect production, the CPC reports influence futures markets of several traded commodities, as reflected by their rapid reaction upon the release of new CPC information (2, 4). Some studies have suggested that the sensitivity of markets to the CPC report has increased over recent years (9).

Despite its information content and long record and its demonstrated capacity to inform decision making, the USDA’s CPC data are rarely used in scientific research. This may be reasonably attributed to the subjective nature of the data, which relies on the personal opinion of the surveyors. A few studies, however, show that CPC data provide valuable information for predicting crop yields (2, 10–15). The crop development data contained in the condition survey have also been used as a means to validate remote-sensing detection of crop progress and phenology at field scales (16–22). The difficulties of dealing with an ordinal variable in the context of quantitative models are certainly another reason for the limited use of this dataset. In the few studies that use CPC data, the researchers employ different transformations to convert ordinal data to another type more amenable to quantitative analysis. Some studies used a weighted sum of the five classes’ percentage values using linear weights between 1 (for excellent condition) and 0 (very poor) to transform the ordinal variable to a continuous one (2, 12). The weights were chosen arbitrarily and were spaced at equal intervals, hence assuming an interval scale for the crop condition item. In other cases multivariate regression techniques were used to determine average yield corresponding to the different condition categories (11, 23). Other authors added the percent-

ages of the two more positive classes (good and excellent) to generate a weekly index of crop goodness status (13–15). These approaches, however, did not solve the main problems with ordinal variables such as the magnitudes of intervals between categories or the precise location of the mean of the scale so positive and negative conditions can be effectively distinguished and quantified.

**Evidence of Spatial and Temporal Effects in the Crop Condition Survey Data**

To date, very few studies have analyzed the existence of systematic temporal and spatial biases in the CPC dataset. Such effects, however, can be reasonably expected in survey data based on the subjective perception of crop health and development. Spatial differences in the perception of crop status could arise from the very different environmental conditions in which crops are cultivated across the country or from technological differences such as the type of cultivars (varieties), cultivation practices, or cropping schedules. Also, the survey data span a relatively long time in which large technological and agronomic, market, and environmental changes have occurred. These effects, however, vary per state. An exploratory plot of the average scores of the survey by state, year, and week within the season reveals several sources of bias in the data (Fig. 2). For context, Fig. 2 also shows auxiliary information such as average yields per state and year and the most prevalent phenological state each week. In Fig. 2 we see noticeable crop status differences between states, and it could be hypothesized that these differences are related to the mean yields obtained in each case, with higher positive CPC scores in states reporting higher yields. There are also major differences between years, with low-yielding years obtaining a higher proportion of low CPC scores. Note that a long-term effect is apparent in the data: Due to technological advances, years that may be considered to have below-average yields at a point in time may have been considered above average during earlier years. The survey data seem to account for these known technology-driven long-term trends in crop productivity. Finally,
and perhaps more surprisingly, there is evidence of biases within the growing season. Negative scores become more frequent as the season advances and the crop reaches critical phenological stages such as silking or doughing. The scores tend to recover slightly at the end of the season when the crop is mature in most fields. The existence of long-term effects as well as differences between start- and end-of-season scores has been documented for corn and soybeans at the national level (14, 24).

Formal Analysis

Our initial preliminary analysis of the crop condition data shows that further and more refined analysis is required to understand the nature of the dataset and maximize its information value. We conducted such analysis using a cumulative link mixed-effects model (25) with a form that allowed exploring the relevant spatial and temporal features of the dataset. The analysis revealed compelling human perception effects in the crop condition survey data not previously reported. However, and more importantly, the analysis resulted in the development of a homogenized, continuous crop condition index that can be used to compare relative crop development and health in space, in time, and within the growing season.

Similarities of the Underlying Models between Crops. Table 1 shows the linked mixed-effects model coefficients of the analysis of crop condition data. The \( \theta \) coefficients are model intercepts that map the original ordinal variable (the crop condition survey classes) into a continuous variable and also provide a precise quantification of the metric distances between the survey categories. Interestingly, these coefficients were almost identical across crops, indicating the existence of an underlying perceptual model shared by all of the surveyors that does not change between crops. In other words, the survey categories (excellent, very poor, etc.) reflect the same degree of anomaly in the status of a crop, independent of the crop type. \( \theta \) categories. Interestingly, these coefficients were almost identical across crops, indicating the existence of an underlying perceptual model shared by all of the surveyors that does not change between crops. In other words, the survey categories (excellent, very poor, etc.) reflect the same degree of anomaly in the status of a crop, independent of the crop type.

Global Long-Term and Intraseasonal Effects. The \( \beta \) coefficients contain the fixed-effect terms of the analysis and have a global effect on the data. The long-term temporal effect (coefficient \( \beta_y \)) was very close to zero in all four crops and was statistically significant only for soybeans and cotton, indicating null or minimal long-term effects. The lack of a long-term trend in the mean of the CPC data is remarkable, considering the persistent increase in yields observed in the four crops during the study period (SI Appendix, Fig. S5). The analysis shows that surveyors account for this effect and adapt their scores to the expected crop performance associated with ever-improving technology and management practices.

Interestingly, the analysis showed the existence of a seasonal effect (coefficient \( \beta_w \)), which was significantly different from zero in all crops except winter wheat. Although the magnitude of this effect is low, as revealed by its SD it is interesting that the coefficients are negative, implying that condition scores tend to develop a low bias as the season develops. There is a logical explanation for this, since most growing seasons begin under normal conditions and more often than not develop normally throughout the season. Adverse events that negatively affect crops, on the other hand, occur less often but rapidly reduce the yield prospects from the normal expectation when the crop

Fig. 2. Mean crop condition reports (Likert plots) for corn per state (Top), year (Middle), and week (Bottom). States highlighted in bold (Top) have available a complete and uninterrupted record from 1987 to 2018. Mean corn yields are shown in the upper axis for each year and for the states that have a complete record over the study period. The most frequent phenological state is also shown for each week, with the following codes: planted (pla), emerged (eme), silking (sil), doughy (dou), dented (den), and mature (mat).

Beguería and Maneta
Table 1. Model coefficients for the four crops: intercepts ($\theta$), fixed effects ($\beta$), SD ($\sigma$), and correlations ($\rho$) of the random effects

| Crop           | $\theta_1$ | $\theta_2$ | $\theta_3$ | $\theta_4$ |
|----------------|------------|------------|------------|------------|
| Corn           | -2.224*    | -1.504*    | -0.523*    | 1.038*     |
| Soybeans       | -2.143*    | -1.357*    | -0.275*    | 1.318*     |
| Cotton         | -2.316*    | -1.487*    | -0.270*    | 1.436*     |
| Winter wheat   | -2.138*    | -1.360*    | -0.263*    | 1.367*     |

|                  | $\sigma_1$ | $\sigma_2$ | $\sigma_3$ | $\sigma_4$ |
|------------------|------------|------------|------------|------------|
| Corn             | 0.354      | 0.197      | 0.197      | 0.550      |
| Soybeans         | 0.198      | 0.163      | 0.169      | 0.124      |
| Cotton           | 0.585      | 0.518      | 0.518      | 0.164      |
| Winter wheat     | 0.361      | 0.478      | 0.478      | 0.396      |

|                  | $\rho_{yw}$ $\rho_{ys}$ $\rho_{sw}$ $\rho_{sy}$ $\rho_{wy}$ $\rho_{ys}$ $\sigma_\rho$ |
|------------------|------------|------------|------------|------------|------------|------------|------------|------------|
| Corn             | -0.477     | 0.044      | 0.346      | 0.496      | 0.044      | 0.346      | 0.496      | 0.044      |
| Soybeans         | 0.034      | -0.058     | 0.034      | 0.034      | -0.058     | 0.034      | 0.034      | -0.058     |
| Cotton           | 0.090      | 0.197      | 0.197      | 0.197      | 0.090      | 0.197      | 0.197      | 0.090      |
| Winter wheat     | 0.948      | 0.044      | 0.044      | 0.044      | 0.948      | 0.044      | 0.044      | 0.948      |

*Significance at the confidence level $\alpha = 0.05$.

was planted. This generates the slightly high bias in the early season ratings and their apparent subsequent decline as the season progresses.

Differences across States. Since the CPC dataset is aggregated at the state level, it is possible to analyze the existence of spatial differences in the model parameters. This can be done by inspecting the random effect coefficients, which capture the variability of each state around the group-level intercept. The coefficient showing the highest variability (after the random term component, to be discussed later) was the state random coefficient ($\sigma_\rho$), which accounts for differences in the mean survey response across states (SI Appendix, Fig. S6). There is a clear relationship between the magnitude of the coefficient and the mean yields obtained in each state. States with high mean yields also had CPC responses with a higher positive bias. Similarly, states with low mean yields tended to have negatively biased CPC responses (SI Appendix, Fig. S7). The implication of this trend is that the perceptual model of the surveyors is homogeneous and spatially invariant: Surveyors could be randomly reassigned from one state to another, and their responses would still be representative of the local crop conditions.

The analysis also revealed differences in the long-term effects associated with each state, with some states showing positive values indicative of a temporal trend toward increasingly more positive survey scores and other states having negative values (SI Appendix, Fig. S8). These differences are related to the different yield trends experienced by each state. States where crop yields increased over time at rates higher than those of the group also tended to have positive coefficients and thus long-term trends in the crop condition survey response. The opposite is also true. This could be confirmed at least for corn and soybeans (SI Appendix, Fig. S9). Finally, our analysis showed that there is a seasonal effect that varies between states on a weekly basis (SI Appendix, Fig. S10) and that these differences are related to the timing of the main phenological stages of the crop, which also vary between states due to agro-climatic differences. This seasonal effect, however, is weaker than the long-term effect (SI Appendix, Fig. S11).

Random Effect. The residual coefficient ($\sigma_\rho$) explained a large fraction of the random variability of the CPC data, as shown by its variance. Once state, long-term, and seasonal effects are accounted for, the corresponding terms of the model, this residual component represents the unbiased crop condition for each state and week. Since it has been formulated as a random effect per state, this component has a zero mean at the state level, with positive values indicating better-than-normal crop conditions (for a given state, year, and week), while negative values would indicate worse-than-normal conditions. For example, Fig. 3 shows the weekly change in the condition of corn, as represented by the random component of the model, in different corn-growing states during 2012. Fig. 3 illustrates how this component reflects the rapid worsening of the crop condition for the states that were affected by the severe drought that started between weeks 25 and 29. The histograms of the random component of the model adjusted for the four crops we consider in this study (Fig. 4) are left skewed, consistent with our previous discussion that the CPC has more categories defining worse-than-normal crop conditions than favorable crop conditions. The random component also displays differences across states (SI Appendix, Fig. S12), and it is interesting to confirm that there is a relationship between the variances of the random component and the yields recorded at each state (SI Appendix, Fig. S5).

Development of a Crop Condition Index

We have shown that, despite the subjective nature of the survey, the CPC data present highly robust characteristics across crops, states, and time. The main obstacle preventing a wider use of the CPC dataset is the ordinal character of the information. Our analysis framework, however, transforms the original data into a continuous variate, more suitable for mathematical analysis. The residual component of the model, once spatial and temporal biases and sources of variability have been eliminated, is therefore proposed as a quantitative crop condition index (CCI) (26). Because this CCI is a continuous variable, it can be used to monitor and assess the status of crops with a higher level of precision. Also, because the CCI is unbiased, it can be used to compare the status of crops between states, between years, and even between crops.

The CCI is, despite being negatively skewed, almost normally distributed; however, because it has different variances between states, it is not a fully standardized index. We have decided to leave the random component as it is and do not do any further transformation to standardize the CCI since these are intrinsic characteristics of the data that need to be preserved. In the next section, we provide an example of how the CCI can be used to provide early prediction of crop yields.

Early Prediction of Crop Yields Based on the Crop Condition Index

Early crop yield forecasts, along with crop acreage published by the USDA, are highly relevant for the agri-food sector. Accurate forecasts of yields are important to inform analysis and decision making. We develop a simple linear model of crop yields at the

---

![Fig. 3. Evolution of the crop condition index for corn during the 2012 growing season at different states. Positive values (in shades of blue) indicate better-than-normal conditions, while negative values (red) indicate worse-than-normal conditions.](image-url)
The yield predictions generated by the CCI-based model are virtually free of bias (mean error), while the USDA forecasts show a slightly higher bias. The CCI-based model also achieves better accuracy than other models that use raw crop condition data (11–13). The CCI-based predictions are similar to and in some cases even closer to the target end-of-season yields. The plots illustrate the advantage of having weekly CCI data, which permits relatively accurate yield predictions many weeks before the first official USDA forecasts are issued. Fig. 6 shows the weekly evolution of goodness-of-fit, error, and bias statistics for the four crops, considering all states, years, and weeks. Color scheme is as in Fig. 3.
The spatial distribution of the goodness of fit is shown in Fig. 7.

Conclusion
Current crop-monitoring methodologies, operational yield forecasts, and early warning systems rely on remote-sensing imagery and crop models and use precipitation, temperature, or other meteorological data to detect anomalies, delineate their extent, and characterize their severity. Meteorological anomalies, such as drought, however, do not always affect agriculture because better management practices and technology often permit growers to maintain production under adverse climatic conditions. Without land surveys, many regions of the world can only rely on remote sensing and meteorological data as the basis for crop monitoring. In the United States, the weekly USDA crop condition survey provides an accurate assessment of the state of crops that integrates all relevant information and biophysical factors and considers specific local technological and management practices, as well as particular circumstances that may affect the timing and regular progress of a crop such as late planting dates. Our results demonstrate that a quantitative crop condition index can be developed based on the qualitative crop condition survey. This index permits the direct comparison of crop conditions across states and years, and its continuous nature makes it more amenable to be used in qualitative research. It also provides superior information that can be used to generate better operational crop monitoring and prediction systems at state scales in the United States.

Materials and Methods

Dataset. We downloaded crop condition data and other auxiliary variables (crop development, yields) from the Quick Stats database of the National Agricultural Statistics Service (28), for four major crops (corn, soybeans, winter wheat, and cotton), for the period between 1987 and 2018. Details on the data availability per state and crop are provided in SI Appendix, Table S2. Crop condition data consist of percentages for each of five condition classes (very poor, poor, fair, good, excellent), aggregated at the state scale, for each week during the growing season of each crop. It constitutes an example of an ordinal (or ordered categorical) variable, such as are widespread in scientific disciplines where humans are utilized as measurement devices, and Likert items are used to get information about a given problem. A Likert item is a simple question for which the response is codified on a discrete ordered scale ranging between two extreme values. The USDA crop condition survey can thus be considered a variety of a Likert item. Ordinal variables contain no metric information since the different levels of response do not indicate equal intervals between them. Therefore, a standard metric analysis is not feasible with ordinal data. There are techniques, however, suited to ordinal variables that allow answering relevant questions such as the distances between categories, the precise location of the mean category, or mean differences between different populations. One of such ways is the cumulative link model with a probit link, also known as the ordinal probit model. The cumulative link model allows transforming an original ordinal variable into a continuous, normal variate described by a mean and a SD. The process involves determining the threshold values that discriminate between the different classes of the variable, as is explained more formally in SI Appendix.

Preliminary Analysis. Likert plots (cumulative bar plots customarily used to portray ordinal variables) were used to explore crop condition survey data stratified per states, years, and weeks and help establish the model hypotheses. The reference line at 0 was set at the middle of the class fair, although further statistical analysis allowed us to determine more rigorously the mean of the distribution of the ordinal variable.

Statistical Analysis. A cumulative link mixed model (CLMM) was used to analyze the crop condition survey data. The model included a long-term linear trend component (variable year) and a seasonal component (variable week) as fixed effects and the state and interactions between state and year and state and week as random components. Another random component was included to represent the random variations that occur each week and on each state. This component reflects the crop condition anomalies, once the effects of state, year, and week have been accounted for. A probit link function was used to relate between these model components and the probabilities of each condition class,

\[
\text{probit}(P(Y_i \leq j | s, y, w)) = \theta_j + \beta y + \beta w + \nu_j + \nu_j y + \nu_j w + \epsilon_i,
\]

where \(P(Y_i \leq j)\) is the probability that condition of record \(i\) would correspond to class \(j\) or lower; \(s, y,\) and \(w\) are the state, year, and week corresponding to record \(i\); \(\theta_j\) is an intercept; \(\beta y\) and \(\beta w\) are model coefficients for the year and week fixed effects; \((\nu_j, \nu_j y, \nu_j w) \sim \mathcal{N}(0, \Sigma)\) are multivariate normally distributed random intercept, year, and week effects; and \(\epsilon_i \sim \mathcal{N}(0, \sigma^2)\) is a random error. The latter term of the model (the random error) represents the crop condition index (CCI) for that particular state, year, and week, once all of the fixed and random effects have been accounted for. The model was fitted for each crop separately, and estimates of the model’s parameters were obtained using a Laplace approximation to the maximum-likelihood function, as implemented in the ordinal R package (29).

Yield Prediction Model. We defined a hierarchical mixed-effects linear model of crop yields as

\[
\mu_s(y) = \beta_0 + \beta_y y + \beta_CCI + \nu(y) + \nu_CCI + \epsilon,
\]

where \(\mu_s(y)\) is the expected yield at state \(s\) and time \(y\); \(\beta_0\) is a global intercept; \(\beta_y\) and \(\beta_CCI\) are model coefficients for the long-term (year)
and CCI fixed effects; \((v_1, v_2, v_3) \sim N(0, \Sigma)\) are multivariate normally distributed random effects; and \(v_i \sim N(0, \sigma_i^2)\) is a random error. The model was fitted for each crop and week during the season separately, and parameter estimates were obtained by the restricted maximum likelihood (REML) method, as implemented in the `lme4` R package (30, 31). P values for the fixed-effects coefficients were computed using the `lmerTest` package (32). Out-of-bag best linear unbiased predictions (BLUPs) were calculated for each crop and year in the dataset, which allowed for an unbiased assessment of the model's predictive power. The 95% prediction intervals around the BLUPs were estimated by drawing a sampling distribution for the random and the fixed effects and then sampling the fitted values across that distribution, as implemented in the `merTools` R package (33).

### Data Availability
The CCI dataset and the R code of the yield prediction model may be downloaded from the institutional repository of CSIC: [http://hdl.handle.net/10261/201950](http://hdl.handle.net/10261/201950).

### ACKNOWLEDGMENTS
S.B.’s research is supported by the Spanish Ministry of Science and Innovation (Grant GCL2017-83866-C3-3-R) and Fundación Biodiversidad, Ministry of Ecological Transition (Grant CA,CC,2018). M.P.M.’s support is from the USDA-National Institute of Food and Agriculture (NIFA) (Grant 2016-67026-25067) and the NASA Established Program to Stimulate Competitive Research (EPSCoR) program (Grant 80NSSC18M0025M). This research was made possible thanks to a mobility grant of the Spanish Ministry of Education, Culture, and Sports, within the framework of the “Programa Estatal de Promoción del Talento y su Empleabilidad en I+D+i,” “Salvador de Madariaga” program.