NVision Acreage & Demand Model

Model Documentation

1. Background

This is a documentation of the modeling workflow in the forecasting of crop acreage and nitrogen demand in the NVision Project. The document covers the data preparation carried out prior to the modeling exercise, the modeling of corn and soybean acreages, and the modeling of total nitrogen and product demand in the US.

The following is a summary of the language and concepts used throughout this document

- Error is measured by the Mean Absolute Deviation (MAD):

$$MAD = (1/\ number\ of\ seasons) * \sum |\ prediction\ value - truth\ value\ |$$

- `[historic_models.py]` The "base model" represents a simple model in counties/districts where more complex models cannot be developed (e.g. the input variables needed are not available). A base model is created by merging (each geographic unit picks only one model) the following historical models based on their out-of-sample errors (in order):
    - Persistence model: Prediction of the current year's value is equal to the previous year's value.
    - Median model: Prediction of the current year's value is the median of the value in the previous five years (script can be generalized for other choices of *N*).
    - Trend model: Assumes a linear trend in the previous five years (script can be generalized for other choices of *N*). Uses a least-squares linear regression fit to predict the current year's value.
- All component models discussed below are some variation of linear regression fit on certain fundamental features. In cases where future improvements can be made, additional procedures such as merging and blending will be carried out to lower model error (see below for more).
    - All features are regularized by the mean and standard deviation for better interpretability in the model coefficients.
- `[merge_models_by_error.py]` The process of "merging" selects the best model in each geography (e.g. county or ag-district). The merged model in each year is selected based on the best out-of-sample errors.
    - For each year and each geography, a script compares all candidate models against the truth value. The process iterates through all candidate models in a prespecified order.
    - In order for a candidate model to be selected, it must outperform the previous candidate model by more than 5% in out-of-sample MAD.

    - As an example, when merging is performed on the 2008 model year, the merged model is selected based on the lowest MAD in all other years except 2008.
- `[blend_models_by_error.py]` The process of "blending" selects the optimal weighting of three component models in each geography (e.g. county or ag-district). The blended model in each year is the weighted average of three component models based on out-of-sample errors.
    - For each year and each geography, a script computes all weighting permutations of the three models in 10% increments (e.g. 60/30/10%, 60/20/20%, etc.). The script then compares each candidate model against the truth value and determines the out-of-sample MAD.
    - In order for a weighting scheme to be selected, it must outperform the previous weighting scheme in the iteration by more than 5% in out-of-sample MAD.
- Backtesting procedure
    - A model for each season for each geography is fit using data from all other seasons for that geography. This ensures no data from the season being modeled is used (i.e. no data leak).
    - However, in performing merges or blends as defined above, there is a small amount of data leak that occurs.
        - For example, if the modeler is performing a merge for a district in 2008, finding the best model involves looking at the out-of-sample errors for all other years; but the out-of-sample errors for other years use information about the error in 2008, and this is a data leak.
        - To ensure that this data leak is not adversely impacting what models are picked in merges or blends, or that errors are misrepresented, the entire modeling pipeline is rerun from beginning to end for all the seasons that are modeled, where for each season any truth data from that season's truth files is removed so there is no chance of leakage.

The following is a summary of the general modeling processes that took place in the development of each NVision acreage and demand model:

1. Create a benchmark (base) model that was representative of standard modeling techniques
2. Develop more complex input features and a model with superior performance relative to the benchmark (base model)
3. Utilize KAES SME knowledge to develop and test additional features into the output model
4. Combine the best models into the blend/merge with customization at the district or county level
    a. Acreage: Merge at the county level
    b. Total Demand / Product Demand: Merge at the district level
5. Validated through extensive backtesting to avoid overfitting

2. Data Preparation

2a. Acreage Data Preparation

- Crop Acreage ground truth source: USDA NASS Crop Acreage (1999-2018).
- `[get_planted_correction.py]` "State-correct" NASS county-level acreage data. This is necessary because when modeling acreage in Nov/Dec of season *X* for season *X+1*, only state-level estimates are available for *X.* By prorating the Descartes Labs-estimated county-level crop acreages to NASS' state estimates, the county-level acreage truth file can be extended one more year (i.e. into 2018).
- `[make_yld_truthfile.py]` Create crop yield truth file (source: USDA NASS).
- `[pull_commodities_prices.py]` Retrieve and process CME corn, soybeans, and winter-wheat futures prices from Quandl API. To be used in crop price model features.
    - The rolling 30-calendar-day mean future prices is calculated for each crop.
    - Contract months by futures:
        - Corn: December
        - Soybeans: November
        - Winter Wheat: July
    - When modeling season Y, the price features used come from the expiration months mentioned in the previous bullet point for year Y.
        - If a 2019 model is being run in November 2018, the December corn contract futures prices will be those from December 2019 expiration.

2b. Nitrogen Demand Data Preparation

- Nitrogen demand ground truth source: AAPFCO Nitrogen Demand Report (1991-2014)
- Demand truth file backfill
    - The state-level demand data is available whereas the county-level data in many states are incomplete (e.g. all county-level data in a state grouped into a single "other county" category). This is a known issue and is caused by the data collection methodology.
    - To ensure that we have the county- and ag-district-level granularity needed to establish the ground truth, a backfill process took place to correct for the missing county-level data:
        - `[get_dist_koch_demand_v11.py]` In states where at least one year of county-level breakdown is available in the AAPFCO report (e.g. 2008-2010 IL county-level data N/A), use the most recent county breakdown and backfill the missing years. This assumes the county breakdown remains constant over the years. The backfilled county-level data is then rolled up to the district level for modeling.

        - `[interpolate_dist_to_dist_v11.py]` In states where no county-level breakdown was available in the modeling period 2000-2014, spatial interpolation is used to backfill missing district-level demand data.
        - Spatial interpolation assumes that the level of demand in each district can be interpolated using the known demand in neighboring districts. The demand for all districts in a state is then rolled up to the state level and prorated so that all districts aggregate to the correct state-level demand figure.
- `[make_district_yld_truthfile.py]` Roll up crop yield data from county to district level for demand modeling.
- `[pull_fertilizer_prices.py]` Process daily NH3, UAN, and UREA prices data received from KAES for fertilizer price model feature.
    - Compute daily average between high and low prices for each product. We then compute the 30-calendar day rolling average prices at the 1st and 15th of every modeling month (November to March each year).
- `[pull_spx_prices.py]` Retrieve and process S&P 500 futures prices from Quandl API.
    - Adjust for roll date convention in quarterly S&P futures contracts.
    - Generates the following S&P features for demand modeling:
        - Log of S&P 500 futures prices, and
        - Delta of log S&P from the moving 252-day average (i.e. difference vs. prior year rolling mean)
- Retrieve and process district-level fall weather data for product demand modeling.
    - `[cfsr_processing.py]` Retrieve the following daily weather variables from NOAA's Global Surface Summary of the Day (GSOD) and Climate Forecast System Reanalysis (CFSR):
        - Daily precipitation (measured in mm)
        - Max, min, and average temperature (℃, to be converted to ℉)
        - Volumetric soil moisture content at the following levels: 0 - 0.1m (used) and 0.1 - 0.4m (not used)
    - `[make_fall-weather.py]` Calculate mean soil moisture when the soil temperature is suitable for NH3 application (between 32 and 50℉).
        - Estimates near-surface soil temp from 3-day trailing mean air temperature
        - Identify days and when 1) soil temperature is between 32 and 50℉, and 2) post-50% harvest progress (source: USDA NASS)
        - Calculates mean soil moisture during “suitable” soil temperature period for product demand modeling
- `[make_shipments_file.py]` Process shipment volume data received from KAES for product demand modeling. The data is collected at the city-level and this script aggregates the data to the state-level. General data scrubbing procedures including the removal of non-US shipment data, date-time conversions, nitrogen tons conversions, etc.

3. Crop Acreage Models

- Build historical corn and soybean acreage models with county-level NASS data from 1999 to 2018. The historical models are Persistence, Median, and Trend. The “base model” is then created by merging historical models based on their out-of-sample errors (see Section 1 for more on base models).
- Crop acreage is modeled using a linear regression fit on the county-level trend and a price feature. Acreage is modeled at the county level.
    - Crop prices represent rolling 30-day mean crop futures prices.
    - In price ratio models, the ratio between the yield-normalized commodities-futures-price ratio of the two crops is used. In each county, the average futures prices for each crop is divided by the past five-year average yield. The price ratio for the first crop is then divided by the price ratio for the second crop.
        - The reasoning is that the acreage for certain crops (especially corn and soy) are very much tied to each other. Thus, any price advantage for one crop over the other should result in changes in planting decisions for farmers.
    - In absolute price models, the yield-normalized commodities-futures-price ratio is only determined for one crop.
- The final corn acreage model is a merge of the following models in order:
    - `[historic_models.py]` In counties or years where there wasn't sufficient data to compute any of the acreage models below, the base model is used
    - `[price_models_v0.py]` The Yield-Normalized Corn-Soybeans price ratio model, and
    - `[price_models_v0.py]` The Yield-Normalized Absolute Corn price model
- The final soybeans acreage model is a merge of the following models in order:
    - `[historic_models.py]` In counties or years where there wasn't sufficient data to compute one of the acreage models below, the base model is used
    - `[price_models_v0.py]` The Yield-Normalized Soybeans-Corn price ratio model, and
    - `[price_models_v0.py]` The Yield-Normalized Soybeans-Winter Wheat price ratio model
- For more on model merging, see Section 1

4. Nitrogen Demand Models

4a. Total Nitrogen Demand

- Build historical total nitrogen demand models with district-level AAPFCO data from 2000 to 2019. The historical models are Persistence, Median, and Trend. The “base model” is then created by merging historical models based on their out-of-sample errors (see Section 1 for more on base models).
- Total demand is modeled using a linear regression fit on a combination of corn acreage and price features (see more below). Total demand is modeled at the ag-district level.
    - In price ratio models, the ratio between two yield-normalized price features is used. In each district, the average of each price feature is divided by the past five-year average yield. The first price ratio is then divided by the second price ratio
        - The reasoning is that the certain price features (e.g. fertilizer and crop prices) are tied to each other. Thus, any price convergence/divergence should result in nitrogen management changes for farmers
- The final Total demand model is a blend of the following models in order:
    - `[historic_models.py]` In districts or years where there wasn't sufficient data to compute any of the demand models below, the base model is used
    - `[demand_corn-acreage_models.py]` Corn Acreage only model
    - `[demand_corn-acreage_models_v1.py]` Corn Acreage + Yield-Normalized Fertilizer-Corn price ratio model
        - Fertilizer price is comprised of the average yield-normalized fall prices for NH3, UAN, and UREA. Fall prices represent the prior year’s fall (i.e. August to December) prices.
    - `[demand_corn-acreage_models_v1a.py]` Corn Acreage + S&P Index model
        - In Iowa districts, SPX represents the log of S&P 500 futures prices.
        - In all other districts, SPX represents the delta of log S&P from the moving prior year mean (252-day average).
        - The reasoning behind the different approaches is that the first SPX feature performed better only in Iowa.
- For the best model accuracy, only the Corn Acreage + Log S&P Index model is used in Iowa (i.e. no blending in Iowa districts).
- For more on model blending, see Section 1

4b. NH3 Product Demand

- Build historical NH3 product demand models with district-level AAPFCO data from 2000 to 2019 (see Section 1 for more on historical models). The Median model is assigned as the NH3 “base model” because it performed better in terms of national level errors than the other historical models.
- NH3 product demand is modeled using a linear regression fit on a combination of corn acreage, total demand, weather, shipment, and price features (see more below). NH3 product demand is modeled at the ag-district level.
    - Price ratios are calculated by dividing the first price feature by the second price feature
        - The reasoning is that the price for certain nitrogen products are tied to each other. Thus, any price convergence/divergence should result in farmers’ decision to pick one nitrogen product over another.
- The final NH3 product demand model is a merge of the following models in order:
    - `[historic_models.py]` In districts or years where there wasn't sufficient data to compute any of the demand models below, the Median base model is used
    - `[nh3_demand_corn-acreage.py]` Corn Acreage only model
    - `[nh3_demand_corn-acreage-weather.py]` Corn Acreage + Fall Weather model
        - Fall Weather represents the mean soil moisture level during the prior year fall that is “suitable” for NH3 application.
        - “Suitable” period is defined as 1) soil temperature is between 32 and 50℉, and 2) post 50% harvest progress (see Section 2b for more on the Fall Weather feature).
    - `[nh3_demand_weather_px-ratio.py]` Total Demand + Fall Weather + NH3/UAN price-per-acre ratio model
        - Fertilizer price-per-acre is defined as:

            $$Product\ price\ per\ acre\ =\ Product\ price\ *Total\ N\ *\ 1\ /\ Corn\ Acres$$

            where Product price is the prices from fall the prior year (i.e. August to December). Prices are not yield-normalized.
        - The reasoning behind this formulation is that this provides an alternative price signal in relation to planted acreage, instead of product weight.
        - The NH3/UAN price-per-acre ratio is the NH3 price-per-acre divided by the UAN price-per-acre.
    - `[nh3_demand_weather_px-ratio_spx.py]` Total Demand + Fall Weather + NH3/UREA price-per-acre ratio + S&P Index models
        - Fertilizer price-per-acre is defined as:

            $$Product\ price\ per\ acre\ =\ Product\ price\ *Total\ N\ *\ 1\ /\ Corn\ Acres$$

where Product price is the prices from fall the prior year (i.e. August to December). Prices are not yield-normalized.
        - The NH3/UREA price-per-acre ratio is the NH3 price-per-acre divided by the UREA price-per-acre.
        - SPX represents the delta of log S&P from the moving prior year mean (252-day average).
    - `[nh3_demand_weather_px-ratio_spx_shipments.py]` Total Demand + Fall Weather + NH3/UREA price-per-acre ratio + S&P Index + Fall NH3 Shipments models
        - Fertilizer price-per-acre is defined as:

$$Product\ price\ per\ acre\ =\ Product\ price\ *Total\ N\ *\ 1\ /\ Corn\ Acres$$

where Product price is the prices from fall the prior year (i.e. August to December). Prices are not yield-normalized.
        - The NH3/UREA price-per-acre ratio is the NH3 price-per-acre divided by the UREA price-per-acre.
        - SPX represents the delta of log S&P from the moving prior year mean (252-day average).
        - Shipment represents the NH3 shipment volume data from KAES. After Jan 1 each year, only the Jan 1 shipment data is used for improved accuracy.
- For more on model merging, see Section 1

4c. UREA Product Demand

- Build historical UREA product demand models with district-level AAPFCO data from 2000 to 2019 (see Section 1 for more on historical models). The Persistence model is assigned as the UREA “base model” because it performed better in terms of national level errors than the other historical models.
- `[historic_models.py]` In districts or years where there wasn't sufficient data to compute any of the demand models below, the Persistence base model is used.
- `[urea_demand_corn-acreage_px-ratio_spx.py]` In districts or years where there wasn't sufficient shipments data, the Total Demand + UREA/UAN price ratio + S&P Index model is used.
    - Fertilizer product prices are the prices from fall the prior year (i.e. August to December). Prices are not yield-normalized.
    - The price ratio is calculated by dividing the average UREA price over the average UAN price.
        - The reasoning is that the price for certain nitrogen products are tied to each other. Thus, any price convergence/divergence should result in farmers’ decision to pick one nitrogen product over another.
    - SPX represents the delta of log S&P from the moving prior year mean (252-day average).

- `[urea_demand_corn-acreage_px-ratio_spx_shipments.py]` UREA product demand is modeled using a linear regression fit on the Total Demand, UREA/UAN price ratio, S&P Index, and Fall NH3 Shipments features. UREA product demand is modeled at the ag-district level.
    - Fertilizer product prices are the prices from fall the prior year (i.e. August to December). Prices are not yield-normalized.
    - The price ratio is calculated by dividing the average UREA price over the average UAN price.
        - The reasoning is that the price for certain nitrogen products are tied to each other. Thus, any price convergence/divergence should result in farmers' decision to pick one nitrogen product over another.
    - SPX represents the delta of log S&P from the moving prior year mean (252-day average).
    - Shipment represents the NH3 shipment volume data from KAES.

4d. UAN Product Demand

- Build historical UAN product demand models with district-level AAPFCO data from 2000 to 2019. The historical models are Persistence, Median, and Trend. The "base model" is then created by merging historical models based on their out-of-sample errors (see Section 1 for more on base models).
- UAN product demand is modeled using a linear regression fit on a combination of corn acreage, total demand, and price features (see more below). UAN product demand is modeled at the ag-district level.
- The final UAN product demand model is a blending of the following models in order:
    - `[historic_models.py]` In districts or years where there wasn't sufficient data to compute any of the demand models below, the base model is used.
    - Total Demand + UAN-price-per-acre model
        - UAN-price-per-acre is defined as:
          $$UAN\ price\ per\ acre\ =\ UAN\ price\ *Total\ N\ *\ 1\ /\ Corn\ Acres$$
          where UAN price is the prices from fall the prior year (i.e. August to December). Prices are not yield-normalized.
        - The reasoning behind this formulation is that this provides an alternative price signal in relation to planted acreage, instead of product weight.
        - After Dec 1 each year, only the Dec 1 UAN prices is used for improved accuracy.
    - Corn Acreage + UAN price model
        - UAN price is the prices from fall the prior year (i.e. August to December). Prices are not yield-normalized.
        - After Jan 1 each year, only the Jan 1 UAN prices is used for improved accuracy.
    - Corn Acreage + S&P Index model

        - SPX represents the delta of log S&P from the moving prior year mean (252-day average).
    - For the best model accuracy, only the Total Demand + UAN-price-per-acre model is used in Iowa and Illinois.
    - For more on model blending, see Section 1.