This section details the provided data variables. Each variable is described and data type is given. Variables are described in the order in which they appear in the data files. Data is courtesy of Ag Analytics.

The fields x and y are used to hold the geographic coordinates of a given record. X holds the data point’s longitude coordinate and Y holds the point’s latitude coordinate. Data are in the form us unprojected, decimal coordinates using the WGS-84 coordinate reference system (CRS). Data are numeric with 8 decimals of precision. This translates to an approximate location precision of 1.11 x 10 ^ -6 m (which is beyond the precision of the RTK-enabled GNSS receiver, so location accuracy, which is more likely accurate to +/- 4 cm).

VRYieldVol is used to store volumetric yield. This is the value corresponding to the yield volume recorded by the harvester’s yield monitor at a given data point. Data are numeric with up to 6 decimals of precision. Units are user-defined and are usually either kg/ha, lbs/ac, or bushels/ac. The data value’s unit is stored in metadata file that is associated with the original harvest data file in John Deere’s data store. It may be possible to retrieve these by accessing the source data in Ag Analytics’s cloud database.

Row and Col store the row and column of the data point’s location within Ag Analytic’s 10m x 10m data point sample grid for a given field. Row and col identifiers are 0-indexed (i.e. column numbering starts with 0 instead of 1). These grid points are aggregations of data falling within the 10m x 10m area represented by a given row and column. Values are calculated from raw data using proprietary algorithms. Data are of the integer data type and should be treated as being categorical in most cases.

CropSeason is generally the year in which the crop used to generate a given data file was grown and harvested. For example, 2008 corresponds to a crop grown in 2008. However, this van be any integer value and is not necessarily a year value. These data are of the integer data type and should be treated as categorical, label values in all analyses.

HarvestId is an internal record identifier corresponding to AnAnalytics’s harvest record id used to retrieve the harvest data (i.e. VRYieldVol aka volumetric yield volume). These data are text and should be treated as categorical in all analyses.

Crop name holds the name of the crop associated with a given record. These values are all text and are generally the common name of a given crop, such as sorghum or corn. These values should be treated as categorical in all analyses.

Relative\_Elevation1 is a decimal value representing the standardized elevation value (z-score) of a given record relative to the mean elevation of the field. This is useful in determining areas of higher and lower elevation present within the field. Relative elevation can affect water and nutrient status of a given area due to how water and nutrients flow to and from a given area due to elevation differences. Values are positive or negative numeric with 9 decimals of precision.

Slope1 holds the maximum slope value present in the 10m x 10m cell represented by a given data point. Values are positive numeric values having 9 decimals of precision. Slope can affect water and nutrient status of a given point due to how water and nutrients flow over that area due to the degree of slope.

TR1 holds the terrain ruggedness index value. This represents the amount of hilliness and slope amount present within a given cell. This value is calculated by a proprietary algorithm. Values are positive numeric with 9 decimals of precision. Terrain ruggedness can affect water and nutrient status of an area by affecting how water and nutrients flow and percolate through a given area.

Elevation1 holds the absolute elevation value of a given point. This is generally meters above sea level (ASL), but is user-configurable on the GNSS sensor. Values are numeric with 8 decimal points of precision.

SeedingDensity holds the seeding rate (# seeds/acre) planted at a given data point (10m x 10m cell). Values are positive integer values. Seeding density can affect plant yields and growth rates. Optimum seeding density varies from crop to crop and with environmental and agroclimatic conditions of a given growing region and season.

SeedingDensity\_missing is a boolean value indicating whether seeding density data are missing. A value of 1 indicates that no seeding density data are available for a given point, while a value of 0 indicates that seeding density data are present.

Seeding\_Variety1 is a categorical variable indicating the variety of a given species that was planted at a given data point. This value can be null, and null values may be indicated by nan in the data files. If a value is present, it will be an alphanumeric value.

Application\_<#>\_<N\_Rate, ID, and date> – these fields hold the application rates of nitrogen (N) applications for each record, the application operations’ associated internal ids and their application dates. There are 10 fields for each of these for a total of 30 total fields. Nitrogen application rate, in conjunction with the percentage of N applied is very strongly correlated to plant growth and yield in most grain crops (soybeans and other leguminous crops being the exception). The N% values can be retrieved from the AgAnalytics database using the associated ID record values from the ApplicationOperations table. N\_Rate values are numeric 8 decimals of precision. Units are user-defined and are generally gallons/ac, lbs/ac, or kg/ha. Application\_ID records are categorical alphanumeric text values and correspond to AnAnalytics’s internal nitrogen application operation record ids present in their database. Application\_Date records are integer values corresponding to the day of the year in which the nitrogen was applied. For example, a value of 2 corresponds to January 2nd of a given year and 365 (or 366 if a leap year) corresponds to 12/31 of a given year. These values can be compared with the PlantingDay1 values (covered later) to determine how many days after planting (DAP) each N application was applied.

Ph\_mean\_30\_60 stores the mean soil pH value present between 30cm and 60 cm for a given 10m x 10m cell and are numeric with 9 decimals of precision. pH is the -log of hydrogen ion activity present in a sample. Values below 7 are acidic and values above 7 are alkaline. Minimum pH is usually 0, with only extremely acidic compounds having pH values below 0. Maximum pH is usually 14, with only extremely alkaline compounds having pH values exceeding 14. Soil pH is generally between 4 and 10, except in extreme cases. Soil pH affects nutrient availability to plants. Optimum ranges are between 5.5 and 6.5 for most agronomic crops. Ranges outside of these values are can have nutrient deficiencies or toxicities.

Clay\_mean\_30, 60, silt\_mean\_30\_60, and sand\_mean\_30\_60 denote the percentages of clay, silt, and sand, respectively, present in the soil between 30cm and 60cm for a given 10m x 10m cell record. The relative percentages of each affect the soil’s texture, nutrient holding capacity, and other chemical and physical soil properties.

Ksat\_mean\_30\_60 denotes the mean saturated hydraulic conductivity of soil between 30cm and 60cm for a given 10m x 10m cell record. These are numeric values having 9 decimals of precision. This value indicates how easily water can percolate through soil once the soil is fully saturated. Higher values indicate greater flow rates, meaning that the soil allows water to flow through it more freely than areas of soil having lower ksat values. Soil’s hydraulic conductivity can influence water and nutrient availability to plants by influencing how long water and nutrients are present in a given area and how quickly mobile nutrients can leach away. Soils having higher percentages of clay generally have lower ksat values where soils having higher percentages of sand generally have higher ksat values.

Om\_Mean\_30\_60 holds the mean organic matter (OM) percentage of the soil for a given 10m x 10m cell record between 30cm and 60cm. Values are numeric with 9 decimal places of precision. OM generally improves soil texture and soil nutrient and water holding capacity, and nutrient availability to plants. OM can also slowly decompose (mineralize) to release nutrients to the soil solution.

GDD2 to GDD11 – these fields hold the growing degree days (GDD) accumulated at measurement period. These values are numeric with 3 decimals of precision. GDD’s are used to estimate the growth stage of a given crop across the growing season. GDD’s accumulate if the mean temperature is above a certain minimum, which varies from crop to crop. The equation for GDD calculation is *GDD = mean temp – base temp if mean temp > base temp.* The theory behind this is that crops will only develop while the temperature is above a certain minimum threshold. Crops generally reach a certain growth stage at specified ranges of GDD accumulation.

Precipitation 2 to Precipitation 11 – these fields hold the cumulative precipitation amounts for a given measurement period for a given 10m x 10m cell record. Values are numeric with two decimals of precision.

Trait 1 to Trait 5 - these fields indicate whether any genetically modified (GM) traits are present in the crop planted in a given record, for example glyphosate resistance. These are text values and should be treated as categorical values in any analyses.

FIPSCode indicates the federal information processing series (FIPS) unique identifier for the record’s geographic area. These values are integers, but should be treated as categorical values in any analyses.

PlantingDay1 indicates the day of year in which the crop was planted. These are integer values from 1 to 365 (or 366 in leap years).

HarvestDay indicates the day of year in which the crop was harvested. These are integer values from 1 to 365 (or 366 in leap years).