Subnational crop statistics of the European Union

STRUCTURE OF THE DATABASE:

The current version of the Subnational crop statistics DB has **13 fields**, including information on:

- REGION -> The code of region based on the Eurostat classification of NUTS 2016.
- CROP_NAME -> The name of crop as in the Eurostat definition.
- YEAR -> Year of data publication.
- VARIABLE -> The indicator measured in the year (A=Area; Y=Yield; P=Production).VALUE -> The value of the indicator reported in the db.
- UoM -> Unit of measure on the variables: Area (ha), Production (t), Yield (t/ha).
- SOURCE -> The data source of value; could be National Institution or Eurostat (Regio DB) or Mixed.
- CALCULATED_R (flagging system) -> If a value has been derived from a NUTS version different from NUTS 2016.
- CALCULATED_C (flagging system) -> If a value has been derived from different crop codes from the ones in use.
- CALCULATED_V (flagging system) -> If a value is missing or null or 0 and has been calculated after the data collection, according to some defined rules.
- ZERO_AS_NULL (flagging system) -> If a value of zero has been turned into a Null because of any inconsistencies.
- COHERENCE_APY (flagging system) -> If there is coherence among values of Area, Production and Yield.
- COHERENCE_CROP (flagging system) -> If there is coherence among values of Total wheat (Total barley), Soft wheat (Winter barley) and Durum wheat (Spring barley).

FLAGGING SYSTEM:

A total of six flags were reported together with the data, representing additional information on data processing and data quality. In general, all the flags are set as Y (i.e. Yes) or left blank, depending on whether the specific condition occurs or not. Fields about coherence are an exception, because for those cases flags are set as Y or N (i.e. Yes or No), depending on whether coherence is verified or not, and left blank when it is not possible to evaluate coherence, due to missing or null values.

In addition, in the field CALCULATED_R, the original NUTS version is provided together with the Y flag, and the field Calculated_V, whenever possible, indicates if the computed value is derived from other indicators or crops aggregation.

IMPLEMENTED RULES:

- CALCULATED_R -> Yes flag if a value has been derived from a NUTS version different from NUTS 2016.
- CALCULATED_C -> Yes flag if a value has been derived from different crop codes from the ones in use.
- CALCULATED_V -> There are several cases where the value can be calculated. In this dataset, whenever possible it is marked Yes if a value can be derived from other indicators or by means of crops aggregation (for Total wheat and Total barley).
    1. Calculate value from existing indicators:

a. A and P exist and have non-null values, Y is missing -> A > 0, Y = P/A
b. P and Y exist and have non-null values, A is missing -> Y > 0, A = P/Y
c. A and Y exist and have non-null values, P is missing -> P = A * Y

d. A, P and Y exist -> Two out of three indicators are zeros and the third one is null, the third indicator becomes zero as well. Example: A=0, P=0, Y=NA then Y=0 and the flag *calculated value* is added on the record correspondent to the Y indicator.
e. A, P and Y exist -> Two out of three indicators are zero and the third one is null, the third indicator becomes zero as well. Example: A=0, P=0, Y=NA then Y=0 and the flag *calculated value* is added on the record correspondent to the Y indicator.
f. A, P and Y exist -> One indicator is null but the remaining two are non-null, then the null indicator is calculated. For calculating Y, A must be positive; for calculating A, both P and Y must be positive.

Note: in the newly calculated record, source is set according to the other indicators (if both are Eurostat/National then Eurostat/National, otherwise Mixed).

2. Calculate value from crops aggregation (for Total wheat and Total barley):

In order to avoid a lack of information in some regions, whenever possible the Total wheat (Eurostat Code = C1100) was derived by aggregating Soft wheat (C1110) and Durum wheat (C1120). Indicators were calculated as follows:

$A_{C1100} = A_{C1110} + A_{C1120}$
$P_{C1100} = P_{C1110} + P_{C1120}$
$Y_{C1100} = P_{C1100} / A_{C1100}$
If P_{c1100} is not available: $Yc1100 = (Y_{c1110} * A_{c1110} + Y_{c1120} * A_{c1120})/A_{c1100}$

1. C1100 is calculated only when is missing (original values are not substituted, even if there is no coherence among crops values).
2. C1100 is calculated if and only if both C1110 and C1120 exist and are not NA (e.g., If only C1110 or C1120 exists, Total wheat is NOT calculated).
3. If both C1110 and C1120 exist, but at least one value is NA, Total wheat is NOT calculated.
4. In presence of null values, for those countries where C1120 is not cultivated, whenever possible $Indicator_{C1100} = Indicator_{C1110}$
5. Only Total wheat is derived (Soft and durum wheat are not derived).

Note: in the newly calculated record, source is set according to the other crops (if both are Eurostat/National then Eurostat/National, otherwise Mixed).

In the same way, whenever possible the Total barley (Eurostat code=C1300) was derived by aggregating Winter barley (C1310) and Spring barley (C1320). Indicators were calculated as follows:

$A_{C1300} = A_{C1310} + A_{C1320}$
$P_{C1300} = P_{C1310} + P_{C1320}$
$Y_{C1300} = P_{C1300} / A_{C1300}$
If P_{c1300} is not available: $Y_{c1300} = (Y_{c1310} * A_{c1310} + Y_{c1320} * A_{c1320})/A_{c1300}$

1. C1300 is calculated only when is missing (original values are not substituted, even if there is no coherence among crops values).

2. C1300 is calculated if and only if both C1310 and C1320 exist and are not NA NA (e.g., If only C1310 or C1320 exists, Total wheat is NOT calculated).

3. If both C1310 and C1320 exist, but at least one value is NA, Total barley is NOT calculated.

4. In presence of null values, for those countries where C1320 is not cultivated, whenever possible $Indicator_{C1300} = Indicator_{C1310}$.

5. Only Total barley is derived (Winter and Spring barley are not derived).

   Note: in the newly calculated record, source is set according to the other crops (if both are Eurostat/National then Eurostat/National, otherwise Mixed).

- ZERO_SET_AS_NULL -> There are several cases and combinations that could occur:
  a. Only A and P are available (A == 0, P > 0) -> A = NA
  b. Only A and P are available (P == 0, A > 0) -> P = NA
  c. Only A and Y are available (A == 0, Y > 0) -> A = NA
  d. Only A and Y are available (Y == 0, A > 0) -> Y = NA
  e. Only Y and P are available (Y == 0, P > 0) -> Y = NA
  f. Only Y and P are available (P == 0, Y > 0) -> P = NA
  g. A, P and Y exist (A == 0, P > 0) and (Y = NA or Y > 0) -> A = NA
  h. A, P and Y exist (A == 0, P > 0) and (Y == 0) -> A = NA and Y = NA
  i. A, P and Y exist (A == 0, P == 0) and (Y > 0) -> A = NA and P = NA

In the following Tables, all the implemented rules for flags CALCULATED_V (from indicators) and ZERO_AS_NULL are summarized:

Table 1: Implemented rules for flagging systems CALCULATED_V (from indicators) and ZERO_AS_NULL.

		YIELD		
AREA	**PRODUCTION**	*0*	*NA*	*Value*
0	*0*	0	0	Value
	NA	0	NA	Value
	Value	NA	NA	Value
NA	*0*	0	NA	Value
	NA	NA	NA	Value
	Value	NA	NA	Value
Value	*0*	NA	NA	Value
	NA	NA	NA	Value
	Value	C	C	Value

		PRODUCTION		
AREA	**YIELD**	*0*	*NA*	*Value*
0	*0*	0	0	Value
	NA	0	NA	Value
	Value	NA	NA	Value
NA	*0*	0	NA	Value
	NA	NA	NA	Value
	Value	NA	NA	Value
Value	*0*	NA	NA	Value
	NA	NA	NA	Value
	Value	C	C	Value

		AREA		
YIELD	**PRODUCTION**	*0*	*NA*	*Value*
0	*0*	0	0	Value
	NA	0	NA	Value
	Value	NA	NA	Value
NA	*0*	0	NA	Value
	NA	NA	NA	Value
	Value	NA	NA	Value
Value	*0*	NA	NA	Value
	NA	NA	NA	Value
	Value	C	C	Value

C = Calculated; NA=Not Available

- COHERENCE_BETWEEN_A_P_Y:
  1. Yes flag:

  a. A, P and Y exist, and have non-null values -> abs(P-(A*Y)) <= 0.01*P.
  2. No flag:

  a. A, P and Y exist, and have non-null values -> abs(P-(A*Y)) > 0.01*P.
  3. Blank:

  a. Coherence cannot be computed because at least one indicator is missing or is null.
  b. All three indicators (A, P, Y) exist, and all have null values.

- COHERENCE_TOTAL_WHEAT:
  1. Yes flag:

  a. For Area or Production: $Indicator_{C1100}$, $Indicator_{C1110}$ and $Indicator_{C1120}$, and have non-null values -> abs ($Indicator_{C1100}$ -($Indicator_{C1110}$ + $Indicator_{C1120}$)) <= 0.01* $Indicator_{C1100}$
  b. For Yield: Y_{C1100}, Y_{C1110} and Y_{C1120}, and have non-null values -> abs (Y_{C1100} - [(Y_{C1110} * A_{C1110} +Y_{C1120} * A_{C1120})/ A_{C1100}]) <= 0.01*Y_{C1100}
  2. No flag:

  a. For Area or Production: $Indicator_{C1100}$, $Indicator_{C1110}$ and $Indicator_{C1120}$, and have non-null values -> abs ($Indicator_{C1100}$ -($Indicator_{C1110}$ + $Indicator_{C1120}$)) > 0.01* $Indicator_{C1100}$
  b. For Yield: Y_{C1100}, Y_{C1110} and Y_{C1120}, and have non-null values -> abs (Y_{C1100} - [(Y_{C1110} * A_{C1110} +Y_{C1120} * A_{C1120})/ A_{C1100}]) > 0.01*Y_{C1100}
  3. Blank:

  a. Coherence cannot be computed because at least one crop value is missing or is null.
  c. All three crop values exist and all have null values.

- COHERENCE_TOTAL_BARLEY:
  1. Yes flag:

  a. For Area or Production: $Indicator_{C1300}$, $Indicator_{C1310}$ and $Indicator_{C3120}$, and have non-null values -> abs ($Indicator_{C1300}$ -($Indicator_{C1310}$ + $Indicator_{C1320}$)) <= 0.01* $Indicator_{C1300}$
  b. For Yield: Y_{C1300}, Y_{C1310} and Y_{C1320}, and have non-null values -> abs (Y_{C1300} - [(Y_{C1310} * A_{C1310} +Y_{C1320} * A_{C1320})/ A_{C1300}]) <= 0.01*Y_{C1300}
  2 No flag:

a. For Area or Production: $Indicator_{C1300}$, $Indicator_{C1310}$ and $Indicator_{C1320}$, and have non-null values -> abs ($Indicator_{C1300}$ -($Indicator_{C1310}$ + $Indicator_{C1320}$)) > 0.01* $Indicator_{C1300}$.

b. For Yield: Y_{C1300}, Y_{C1310} and Y_{C1320}, and have non-null values -> abs (Y_{C1300} - [(Y_{C1310} * A_{C1310} +Y_{C1320} * A_{C1320})/ A_{C1300}]) > 0.01*Y_{C1300}.

3 Blank:

a. Coherence cannot be computed because at least one crop value is missing or is null.

c. All three crop values exist and all have null values.