Coffee Rust Data
Updated 27 November 2018 | Laura Kahn

Problem: Predict daily coffee rust (Arábica variety) at the country level with machine learning.

Data:

- 1584 Weekly Observations
- From various coffee-growing regions of Brasil
- January 1, 1991 – July 30, 2018 (27 years)

Input variables

- Date (MM/DD/YYYY)
- Temperature (°C)
- Rain (mm)
- Production (1000-60 kg bags of coffee arábica beans
- Futures (USD per pound)

Target variable

Coffee Rust (percent rust disease on leaf)

Note
* Google Translate was used for documents in Portuguese and Spanish

DATA ACQUISITION

- Rust data was read manually off PDF figures and tables existing in the literature. There was no known digital rust data. Data collection methodology affects accuracy of predictions.
- Temperature, rain, production and futures data was all available electronically in CSV format from three different sources.

Coffee Rust Data
Updated 27 November 2018 | Laura Kahn

VARIABLES

Rust

Frequency: Monthly
Precision: Data rounded to nearest whole number since lacking more granularity
Scope: Limited to country-level (not specific coffee growing region) data

1991-1995: Chalfoun Figure 1
1998-2004: Japiassu Figure 1 (low density fodder soil)
October 2005-August 2006: Lopes figure 1 (organic)
December 2007 – November 2008: Lopes figure 1 (organic)
January 2018-July 2018: Matiello Tables 1 and 2

Temperature, Rainfall

Frequency: Monthly (same value used for each weekly observation)
Precision: Data measured to four decimal places
Scope: Temperature & rainfall is average monthly for entire country, not only coffee-growing regions

1991-2015: World Bank Climate Change Knowledge Portal
2016-2017: Used historical monthly data from 2015 since no other data available
2018: Rainfall – Matiello Tables 1 and 2; Temperature– used historical data from 2015 (request 11/10 from World Bank no reply)

Production

Frequency: Yearly (Monthly calculated by dividing yearly total by 12)
Precision: Rounded to two decimal places
Scope: Statistics run from April to April each year since coffee is harvested in April in Brasil
Statistics include both arábica and robusto varieties. No other data available.

1991-2018: International Coffee Organization (Supply data)

Futures

Frequency: Data available daily but used weekly instead
Precision: Rounded to at least three decimal places
Scope: Closing price
Values not modified for inflation

1991 -2018: https://www.macrotrends.net/2535/coffee-prices-historical-chart-data

Other references: http://principo.org/pdi--plano-de-desenvolvimento-institucional-escola-agrotcnica.html?page=17

Coffee Rust Data
Updated 27 November 2018 | Laura Kahn

FEATURE ENGINEERING

- General rule of thumb is to have 10 times number of features to have “enough data”.
- Original dataset from November 2017 had 337 observations from Brasil, Colombia and Papua New Guinea from 1995-1996, 2002-2005 and 2006 with the same input and target variables described above.
- New datasets described above were found in November 2018.
- Production data is only available yearly so in order to create a bigger dataset, I added a weekly observation for each month in the dataset. This added 972 observations (27 years * 12 months * at least 3 extra weekly observations per month).
- Other weather data such as humidity and soil conditions are too sparse in the literature to include in the analysis at this time. Rainfall and temperature were selected based on prior research as being the top two most important weather variables that may affect rust, production and prices.

MISSING DATA

Dataset has 773 / 1584 missing Rust values (48.8%)

Values considered missing completely at random since no data found from those years. See https://en.wikipedia.org/wiki/Missing_data for more information.

Curated dataset has 109 / 1584 Rust Values > 50%

kNN Data imputation method described by Gustavo Batista et al. is used for missing values

Final data: