6th International Congress on Scientific Research August 18-20. 2023, Ankara by IKSAD Institute

**Simulator development for yield estimation (in case of corn, oats, soybean) based on weather-data**

Dániel Váradi (<https://orcid.org/0000-0001-9610-8566>)

Dr. László Pitlik (<https://orcid.org/0000-0001-5819-0319>),

Marcell Pitlik (<https://orcid.org/0000-0002-8058-9577>)

e-Mails: [danielvaradi140@gmail.com](mailto:danielvaradi140@gmail.com), [pitlik@my-x.hu](mailto:pitlik@my-x.hu), [pitlikmarci@gmail.com](mailto:pitlikmarci@gmail.com)

Kodolányi János University and MY-X research team Hungary

*Keywords: forecasting, antidiscrimination, weather, yield, production function, consistence-oriented modelling*

**Abstract**

There are given information units about yields and weather constellations. The yields and weather data are available for well-known regions for more than 100 years. Yields are annual data, the weather is descripted in daily rhythm (through average temperature, minimum, maximum temperature, and precipitation). All time-series have different lacks: e.g., not all the plants were produced in all regions in all years and/or weather data could not be measured for each day - therefore the daily weather data are in general not given for all the days. This situation should be accepted as a real data management situation. The region (US-state) is Minnesota with a lot of counties as production areas. Weather stations are given more than production areas, and the GPS-coordinates of production areas and weather stations present arbitrary differences.

**The task is simple:** a simulator should be developed where yields of the plants (at first corn, oats, soybean) can be derived based on weather data. The developer is free: the developer may decide about the model inputs (based on the prepared data assets: <https://miau.my-x.hu/miau/298/quant>). The test of the simulator is also simple: there are 198 regions (without GPS co-ordinates but with differently completed weather data, it means for different years and with different lack in the time-series). The simulator should deliver yields for the unknown regions based only on their weather data for each year where at least one single weather data is given.

**Solution:** It is possible, based on antidiscrimination models of the own similarity analysis tool (<https://miau.my-x.hu/myx-free/>) to derive the most similar (well-known) region (with well-known soil/environment-reactions) for each unknown region based on the two sets of different weather data. The similarity analysis is capable of searching for the most similar weather situation based on daily and/or aggregated differences in form of an online Solver-oriented approach. If the most similar region is given, then its weather data and yields can be used for region-specific simulator development.

The region-specific simulator development is also a similarity-driven approach where production functions (with staircase functions in the core system) are estimated: e.g., monthly averages for average temperature, minimum, maximum temperature and precipitation can be derived based on the available data for each year (covering the partial data lack). It means averages (as Xi) are responsible for the US-state Minnesota as such. The average yield (at first unweighted) can also be derived in frame of a pivot table as a kind of Y values for the modelling.

These average yields demonstrate a massive increasing trend between 1950-2021. There are data available before 1950 (source USDA: quick stats – with data from 1921). The linear trends for different plants (corn, oats, and soybean) can be estimated with R2-values over 0.9. Polynomials (3th) (it means S-curves) deliver higher R2-values – concerning the expectations. The nominal yields can be transformed into weather-impact-values if the difference is calculated between the raw yield and the trend-based yield (year-by-year). The simulators for raw yields (where the inputs are the monthly weather data of the previous year) are error-free in the learning phase but the test year (2022) presents a potential yield between 0 and the genetic potential. It means: these models are still not useful. If the dependent variable is not the nominal (raw) yield (in BU/acre) but the transformed yield (see weather-impact - based on the linear trends) and let alone not for the same year but for the next one, then the simulators are also error-free – and these models/simulators are capable of real estimations (forecasts) for corn/oats/soybean.

**Conclusion:** the corn for the entire Minnesota should deliver a yield in case of 2022 over the well-known maximum yield before! The soyabean should deliver a yield better than in 2021. Oats yield should also be higher than in 2021. These estimations could be checked based on available data by USDA (e.g., <https://quickstats.nass.usda.gov/>). All these estimations can be evaluated as correct! It means this kind of simulator development (with real forecasting characteristics) is robust enough for entire Minnesota and the logic can be adapted for arbitrary (known and unknown=estimated) regions (counties) based on the antidiscrimination pre-analysis in case of each seemingly unknown test region. The automation of the already tested and robust development-process is trivial if the decision makers give a sign of willingness.

The above-descripted approach is a kind of consistence-driven method for developing simulators (production functions). Consistence means the partial results can be interpreted compared to each other as useful/appropriate puzzle pieces. If these puzzle pieces cannot join with each other, the simulators will not gossip like chatGPT would do in such a case. Parallel, the similarity analyses have an internal quality assurance layer: the mirrored inputs should deliver mirrored outputs. If these double checking is not given, then the simulators will not gossip for affected years and/or regions.