Causal Artificial Intelligence Models of Food Quality Data

Running head: Causal AI Food Quality

Želimir Kurtanjek*

University of Zagreb Faculty of Food Technology and Biotechnology, Pierrotijeva 6, 10000 Zagreb, Croatia

Received: 19 July 2023
Accepted: 27 December 2023

SUMMARY

Research background. The motivation of this study is to emphasize the importance of artificial intelligence (AI) and causality modelling of food quality and analysis with “big data”. AI with structural causal modelling (SCM), based on Bayes networks and deep learning, enables the integration of theoretical field knowledge in food technology with process production, physical-chemical analytics, and consumer organoleptic assessments. Food products have complex nature and data are highly dimensional, with intricate interrelations (correlations) and are difficult to relate to consumer sensory perception of food quality. Standard regression modelling techniques such as multiple ordinary least squares (OLS) and partial least squares (PLS) are effectively applied for the prediction by linear interpolations of observed data under cross-sectional stationary conditions. Upgrading linear regression models by machine learning (ML) accounts for nonlinear relations and reveals functional patterns, but is prone to confounding and fails predictions under unobserved nonstationary conditions. Confounding of data variables is the main obstacle to applications of the regression models in food innovations under previously untrained conditions. Hence, this manuscript focuses on applying causal graphical models with Bayes networks to infer causal relationships and intervention effects between process variables and consumer sensory assessment of food quality.

*E-mail: zelimir.kurtanjek@gmail.com
Experimental approach. This study is based on the literature available data on the process of wheat bread baking quality, consumer sensory quality assessments of fermented milk products, and professional wine tasting data. The data for wheat baking quality are regularized by the least absolute shrinkage and selection operator (LASSO elastic net). Applied is Bayes statistics for evaluation of the model joint probability function for inferring the network structure and parameters. The obtained SCM models are presented as directed acyclic graphs (DAG). D-separation criteria is applied to block confounding effects in estimating direct and total causal effects of process variables and consumer perception on food quality. Probability distributions of causal effects of the intervention of individual process variables on quality are presented as partial dependency plots determined by Bayes neural networks. In the case of wine quality causality, the total causal effects determined by SCM models are positively validated by the double machine learning (DML) algorithm.

Results and conclusions. Analysed is the data set of 45 continuous variables corresponding to different chemical, physical and biochemical variables of wheat properties from seven Croatian cultivars during two years of controlled cultivation. LASSO regularization of the data set yielded the ten key predictors, accounting for 98 % variance of the baking quality data. Based on the key variables derived is the quality predictive random forest model with 75 % cross-validation accuracy. Causal analysis between the quality and key predictors is based on the Bayes model depicted as a DAG graph. Protein content shows the most important direct causal effect with the corresponding path coefficient of 0.71, and THMW (total high molecular glutenin subunits) content is an indirect cause with a path coefficient of 0.42, and protein total average causal effect (ACE) is 0.65. The large data set of quality fermented milk products includes binary consumer sensory data (taste, odour, turbidity), continuous physical variables (temperature, fat, pH, colour), and three grade classes of consumer quality assessment. Derived is a random forest model for the prediction of the quality classification with an “out of box” (OOB) error of 0.28 %. The Bayes network model predicts that the direct causes of the taste classification are temperature, colour, and fat content, while the direct causes for the quality classification are temperature, turbidity, odour, and fat content. Estimated are the key quality grade average causal effects (ACE) of temperature -0.04 grade/°C and 0.3 quality grade/fat content. The temperature ACE dependency shows a nonlinear type as negative saturation with the “breaking” point at 60 °C, while for fat ACE has a positive linear trend. Causal quality analysis of red and white wine is based on the large data set of eleven continuous variables of physical and chemical properties and quality assessments classified in ten classes, from 1 to 10. Each classification is obtained in triplicates by a panel of professional wine tasters. A non-structural double machine learning algorithm
(DML) is applied for total ACE quality assessment. The alcohol content of red and white wine has the key positive ACE relative factor of 0.35 quality/alcohol, while volatile acidity has the key negative ACE −0.2 quality/acidity. The obtained ACE predictions by the unstructured DML algorithm are in close agreement with the ACE obtained by the structural SCM models.

**Novelty and scientific contribution.** Presented are novel methodologies and results for the application of causal artificial intelligence models in the analysis of consumer assessment of the quality of food products. The application of Bayes network structural causal models (SCM) enables the d-separation of pronounced effects of confounding between parameters in noncausal regression models. Based on SCM, inference of average causal effects (ACE) provides substantiated and validated research hypotheses for new products and support for decisions of potential interventions for improvement in product design, new process introduction, process control, management, and marketing.

**Keywords:** Bayes network; AI causality; intervention effects; ACE, food quality

**INTRODUCTION**

According to the EU Commission report by Knowledge Centre for Food Fraud and Quality (KC-FFQ) based on 30 000 consumer responders, perceived food quality as “very important” and contribute 65 % in priority decisions in consumption, compared to food price with 55 % of importance (1). The concept of consumer-perceived food quality is very complex, it is an untaggable interaction of the objective measurable physical-chemical properties and numerous subjective factors such as consumer population culture, ethical issues, economic and social status, tradition, personal preferences, and expected nutritional benefits. It is a multi-dimensional concept which is influenced by a wide range of unmeasurable situational and contextual factors. To food producers, these complexities are difficult to rationalize for possible applications of statistical and mathematical decision-making algorithms. The objective characterization of food complexity can be greatly rationalized using “big data” generated with high throughput analytical instrumentation. Automation of instrument-measurable sensory attributes led to the development of systems such as electronic noses, e-tongue, near-infrared spectroscopy (NIR), infrared spectroscopy (IR), photo-acoustic detectors, and computer vision (2-6). They are applied for online production monitoring, process control, and food safety. The fusion of physical-chemical and electronic sensory data with computer vision enables the application of machine learning models for the detection of specific signal patterns.
It helps food companies in recognizing patterns which drive consumer choice of specific products and improve the chances of continued purchase and potentially in the innovation of new products optimally adjusted to specific markets. Commonly are applied statistical models such as principal component analysis (PCA) and partial least square regression (PLS) and advanced machine learning (ML) algorithms such as artificial neural networks (ANN), convolution neural networks (CNN), decision trees (DT), and random forests (RF). They analyse large data sets of food quality parameters such as appearance, texture, taste, and odour, and identify patterns that may be difficult for humans to detect. Importantly, they can help in identifying food contamination, spoilage, and adulteration, which are crucial factors in maintaining food safety. The main benefit of ML models is the ability to provide an “in-time” assessment of the statistically significant status of food products (7-10). Integration of ML models with business knowledge in a food company on a production system level leads to industrial artificial intelligence (AI). It collaborates by supporting and enhancing of the human thinking process, enables knowledge management and storing, and most importantly, it can learn new knowledge. A bibliometric decade study since 2012 on AI related to food science and technology shows an exponential increase (11). Literature reports indicate that besides academic research there is also very strong interest in AI in major companies in the food industry. Dominant industry interest is in the application of intelligent robotics in specific process unit operations and their integration into a whole company AI-supported management. Besides standard engineering applications, AI is becoming a key support in the discovery and introduction of food innovations such as new components for taste, flavours, and fragrances, especially aimed to reduce the content of sugar and salt in foods and beverages (12). Recent advancements are focused on the integration of food company policymakers, business intelligence and industrial production AI systems (13-16). The success of AI global integration in food-producing companies depends on understanding the human subjective component of present and assumed potentially new markets (Fig. 1). Understanding of intricate dependencies of human subjective and the objective physical-chemical data requires higher levels of AI models on the scheme of “knowledge ladder” (17). Most of the present AI models fit the first knowledge ladder rung with potentially high flexibility and prediction accuracy under unchangeable model training conditions. In a food company, policy decisions of business management and production technologies need AI’s ability to act under new model untrained conditions. Most unforeseen new conditions are due to disruptions of supply chains, effects of climate change on the production of agricultural raw materials, competitor products, and shifts in market preferences. The key upgrade of the present AI models is the application of causal data fusion of human subjective and instrumental objective data. Causal AI
models are on the second and third rungs of the knowledge ladder. Causal relations are deduced from field knowledge (economy, engineering, physics, chemistry, nutrition) and deduced from “big data” by statistical and knowledge models. The causal relations are integrated into AI models as Bayes networks. They enable causal analysis elimination (blocking) of numerous confounding relations present in integrated “big data” training sets. On the second knowledge rung causal AI models are applied for predicting effects (ACE, average causal effect) of potential policy decisions and/or production interventions, labelled as doing operator do(x) in Fig. 1. On the third rung causal AI models are applied for counterfactual reasoning approaching human imaginative intelligence (17).

The aim of this manuscript is to apply causal AI modelling of food quality assessed by consumers and a professional panel of evaluators of wheat baking quality, milk fermented products and wine quality.

MATERIALS AND METHODS

Wheat quality

Analysed is the baking quality of seven wheat winter cultivars from the Slavonia region in eastern Croatia. As the baking quality test applied is the volume of bread loaf under the standard baking protocol. The cultivars were grown for a period of three years under controlled conditions at the experimental field of the Agricultural Institute Osijek, Croatia. Their quality properties were evaluated by 45 physical, chemical, and biochemical variables. Each parameter is determined in triplicate during three consecutive years of cultivation. The measured variables were grouped as 6 indirect quality parameters, 7 farinographic parameters, 5 extensographic parameters, and 25 reversed phase-high performance liquid chromatography (RP-HPLC) of gluten proteins. The experiment methodology and the data are available in the published manuscripts (18,19). All properties are stored as a table of continuous numerical variables. The data are highly correlated and the average absolute Pearson correlation is R=0.41. Principal component analysis of the total data set reveals that the cumulative effect in explaining the total data variance by the first three components is 76.45 % and the first four components account for 82.68 %.

Dairy quality

This dairy dataset contains 1059 samples of consumer quality assessments of fermented milk products (20,21). The dataset consists of 7 variables: pH, temperature, taste, odour, fat, turbidity, and colour. Temperature, pH and colour are instrument-measured properties defined as continuous variables. The average and standard deviations for pH is 6.63+/−1.4, milk pre-treatment temperatures are in the range from 34 to 90 °C, with an average temperature of 44.2 °C. The colour data are
determined spectroscopically with low variability of 1.7 % relative standard deviation. The samples of the physical variables have a non-Gaussian probability distribution. Spearman rank-order correlation coefficients between temperature and pH, colour and odour are significant with an average value of rho=0.25, while the rho correlation between colour and odour is insignificant. The consumer quality evaluation is the ordinal categorical variable with three levels: low, medium, and high. Spearman rank-order quality grade correlation with temperature, colour and odour is significant, while pH is insignificant.

**Wine quality**

The wine quality is a large dataset, 1599 red and 4898 white samples of the Portuguese “Vinho Verde” variants, characterized by 12 physical and chemical composition data and quality assessments provided by a panel of professional wine tasters (22-24). The data file is available from the UCI Machine Learning Repository from the University of California at Irvine. The variables are fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulphur dioxide, total sulphur dioxide, density, pH, sulphates and alcohol. The wine compositions are continuous numerical variables and the quality is an ordinal categorical variable with levels 1-10. The variable density is removed from the data set due to its very high VIF (variance inflation factor) since it is a common effect (causal collider) and hence cofounds modelling parameters (24). The probability distributions of the variables are approximately Gaussian. The data are highly correlated and the first three principal components for the red and white wines account for 99.7 and 99.8 % of the respective variances. Both red and white wines have citric acid maximum relative data variability of 71 %, given as the ratio of standard deviation and mean. The maximum Pearson correlations of the quality are with the content of alcohol, R=0.48 and R=0.44 for the red and white wines. The maximum negative correlation is with volatile acidity, R=-0.39 and R=-0.19 for the red and white wines.

**Methodology**

The basic principles of causal AI modelling are based on the concepts of Bayes statistics and networks (BN). Bayes statistics account for prior knowledge (old model) with new experimental observations (data) in the prediction of a new model. The nature of prior knowledge accounted for in modelling includes deductive (known theoretical knowledge) and inductive (empirical structures and model parameters known from previous studies). Knowledge of a causal AI model is expressed as a joint probability density function \( P \) of the model conditioned on new data. Causal AI modelling is a
two-stage process in which the first objective is to determine the structure of a BN graph \( G \), and in the second stage to determine functional causal dependencies between variables followed by estimation of the model parameters \( \theta \).

\[
P(\text{model}|\text{data}) = P(G=\text{graph}, \theta=\text{parameter} | X=\text{data})
\]

The two-stage process of structural causal modelling (SCM) is expressed as a product of the corresponding probability density functions:

\[
P(G, \theta | X) = P(G | X) P(\theta | G, X)
\]

With inferred causal structure \( G \) and parameters \( \theta \) model posteriori distribution is expressed by the basic Bayes relationship:

\[
\text{Posteriori}(\text{model}|\text{data}) = \frac{\text{Likelihood}(\text{data}|\text{model}) P(\text{model})}{\text{Evidence}(\text{data})}
\]

In case of a model with continuous random variables (Gaussian) it is explicitly expressed in a functional form as:

\[
\pi(\theta | X) = \frac{L(X|\theta)\pi(\theta)}{\int_\theta L(X|\theta)\pi(\theta)d\theta}
\]

For statistical inferences from the model multivariable posterior probability distribution \( \pi(\theta | X) \) applied is extensive sampling by Monte Carlo Markov chain (MCMC) algorithm.

Commonly, the basic modelling presumes is that all considered causal effects are directional, \textit{i.e.}, recurrent causal effects are not considered. It results in model graphs without close loops, and consequently are named as directed acyclic graphs (DAG). Due to DAG Markovian properties \( w \) greatly simplify modelling of complex multivariable random systems \( (25) \). Complete causal DAG graph \( G \) is a set vertices \( V \) (corresponding to the model random variables \( x_i \)) connected with a set \( E \) of oriented edges (arrows), \( G = (V, E) \). It is a Bayesian network (BN) with Markovian property enabling decomposition of a joint probability density function \( P \) as a product of individual nodes (variables \( x_k \)) probabilities \( p \) conditioned on their parent variables \( Pa \). The parent variables are those variables \( x_i \) (vertices) pointing directly to \( x_k \) via a single edge.

\[
P(X) = \prod_{k \in V} p(x_k | Pa(x_k))
\]

Causal dependencies, direct and total, depend on set of network paths between a cause-and-effect variables. To infer causality confounding of interfering variables must be blocked by directed \( d-\)
separation which implies conditional independence in the probability distribution (17). Variables which block interfering interactions define adjustment sets enabling deconfounded (linear and/or nonlinear) estimation of average causal effect (ACE). For models with continuous variables ACE is evaluated as the derivative of expected value of output variable (effect) Y with respect to change of input (cause) X at constant covariates, named as intervention of cause by doing do(x) (6). In case of a linear SCM model, ACE is a value corresponding to average change of effect Y due to intervention by changing cause X for a unit value. For nonlinear SCM models ACE is a function of the cause X defined by the partial derivative:

\[
ACE(Y_x) = \frac{\partial}{\partial x} E(Y|do(x))
\]

RESULTS AND DISCUSSION

Wheat baking quality

The wheat data are regularized by the application of a flexible net of least absolute shrinkage and selection operator (LASSO) as a combination of L1 and L2 norm penalty functions (26):

\[
Lasso(\lambda, \gamma) = \min_\beta \left( \frac{1}{N} \sum_i (y_i - \beta_0 - X^T \beta)^2 + \sum_l |\beta_l| + \gamma \sum_l \beta_l^2 \right)
\]

The initial space of 45 wheat chemical, physical and biochemical variables is reduced to the space of 10 features obtained by optimisation algorithm provided with “glmnet” software (27,29). The optimal selected features are: protein, wet gluten, falling number, water absorption, dough resistance, ratio resistance/extensibility, total glutenin, total high molecular weight glutenin, alpha gliadin, and dough degree of softening.

The model is the assembly of 500 trees obtained with at each random split of 3 variables. Validation of the prediction model shows that with the untrained “out of box” samples it accounts for 75 % of variance (30). Performance of the model predictions is depicted in Fig. 2. Causal relations between the key variables are evaluated as a directed acyclic graph (DAG). The DAG is depicted with the key variables as the nodes, associations between the variables as the edges, and the causal dependences as arrows. In the process of causal structural learning the graph edges and orientations of arrows are considered as random variables with statistical properties estimated by Monte Carlo Markov Chain (MCMC) sampling from Bayesian posteriori distribution, provided as BNDAG software support (31,32). The result is structural causal model (SCM) as graph depicted in Fig. 3. The causal strengths, with positive and negative effects, are given as the path coefficients, which are calculated...
from corresponding d-separated (directionally separated subgraph) adjustments sets by ordinary least squares (OLS) regression with normalized data \((17,33)\).

The SCM model causal inferences are compared (validated) by unstructured causal model with double machine learning (DML) algorithm for estimation of the average causal effect (ACE) \((34)\). The effects are estimated as the ratio of covariance and variance of the residuals of volume \(V\) and \(k\)-th variable \(x_k\) predicted by corresponding random forest mode (RF) \(l\):

\[
ACE(x_k, V) = \frac{\text{cov}(x_k - RF(x_k | X_{-k}), V - RF(V | X_{-k}))}{\sigma^2(x_k - RF(x_k | X_{-k}))}
\]

The ACE estimates with standardized data are presented as a bar chart (Fig. 4). The SCM model and the ACE estimates confirm dominant positive effects on bread baking quality \((V)\) of protein \((P)\) and total high molecular weight (THMW) content.

The main technological benefit is application of the structural causal model (SCM) ability to predict unconfounded effects of intervention action, i.e. “doing effects” \((17)\). Applied is do\((x)\) operator to redesign original DAG and accordingly modify the joint probability function by replacement of random variable \(X_k\) with preselected deterministic value \(x_k\) and d-separation of confounding variables which simultaneously interfere with the intervention (treatment) and effect (outcome). To account for nonlinearity and probability in uncertainty of do\((x)\) effects developed are Bayesian Neural Networks (BNN) \((35)\). The intervention effects of the key causal variables \(P\) and \(THMW\) on bread baking quality \(V\) are presented in Fig. 5. The distributions of the effect \(V\) indicate considerable uncertainty due to the covariates from the adjustment sets, and modest saturation type nonlinearities.

**Dairy product quality**

Causal analysis of the dairy product quality data is based on SCM model. Causal structure network is learned by hill-climbing (HC) algorithm of greedy search of DAG space of association structures and causal directions to optimize Bayes Information Criteria (BIC) \((36)\). Obtained is a relatively simple DAG network depicted in Fig. 6. Temperature and fat content are identified as the exogeneous variables which with product quality and taste is common effects as colliders. The endogenous variables are product pH, odour, turbidity and colour. The product taste and quality grade have common causal ancestors with maximum negative correlation between the grade and temperature \(R=-0.45\) and maximum positive correlation between taste and fat content \(R=0.32\).
Predictive power of the random forest model with 500 trees and 2 randomly selected variables at each is very high yielding with “out of box” average classification error <1 % (30). The maximal causal effect on quality as negative ACE on temperature is -0.04 quality grade/°C in the temperature range 25–60 °C. The fat content ACE on quality grade is 0.4. Functional dependences of ACE are obtained by the adjusted d-separated variables by Bayes Neural Network presented as partial dependent plots in Fig. 7. Temperature ACE is highly nonlinear with the saturation low point at about 60 °C, while fat content ACE is positive and linear in the full range.

**Wine quality**

For the wine quality detailed description of structural modelling (SCM) and causal analysis is given in (24). Here are the causal effects determined by SCM validated by unstructured double machine learning DML causality model (34). Applied is the same procedure is given in Eq. 8. The random forest modelling is applied to standardized data sets separately for red and white wines. Obtained are the models with relative average prediction errors of 5.13 and 4.17 % for red and white wines. The comparative red and white wines ACE are jointly presented in Fig. 8. Alcohol content, predicted by DML and SCM models, has the highest positive ACE on quality of red and white wines. Content of sulphates and free sulphur dioxide have the second most important positive ACE for both red and white wines while volatile acidity has the highest negative ACE. Although SCM and DML models are based on different assumptions corresponding ACE estimates are qualitatively and numerically approximately in agreement.

**CONCLUSIONS**

This manuscript provides methodologies of causal AI modelling applied to complex problem of integration of objective (instrumental) and subjective (human) food quality data. The obtained causal network model enables food engineers with intervention decisions for existing and innovation of new technologies. The methodologies are illustrated by the models of bread baking quality, fermented milk products and wine.

Applied are machine learning models of neural networks and random forest of decision trees. The key research objective is discovery of the causal relations between objective physical-chemical data and human consumer perception of quality. For discovery of causal relations between complex data of wheat biochemical and physical properties and bread baking quality applied is Bayesian statistical model with MCMC sampling of the posterior distribution. Structural causal learning and
analysis of dairy products is achieved by hill-climbing optimization of the network BIC (Bayes information criteria). Besides the structural causal models, for the wine quality data applied is the unstructured algorithm of double machine learning (DML) models with the random forest decision trees.

The main technological application of the presented causal artificial models is ability to evaluate effects of interventions ("doing", do(x) operator) as improvements in production processes parameters and compositions of food ingredients. The causal models enable discovery of process control patterns and support technological decisions outside available regression data. Here, for each presented model evaluated are average causal effects (ACE) based on d-separation criteria and selection of the corresponding unconfounding adjustment sets. For the models of wine quality comparison, the structure based ACE are in agreement with the estimates by unstructured double machine learning (DML) algorithm. Importance of nonlinear causal effects are modelled by Bayes neural networks with d-separated minimal adjustment sets and depicted as partial dependency plots.

CONFLICT OF INTEREST

This author declares that there is no conflict of interest and this research has not received financial support.

AUTHOR CONTRIBUTION

This author is the only contributor.

ORCID

Ž. Kurtanjek https://orcid.org/0000-0001-5453-6255

REFERENCES


https://doi.org/10.3389/fbioe.2016.00081


https://doi.org/10.1016/j.aiia.2020.06.003


https://doi.org/10.3389/fnut.2022.1074958


https://doi.org/10.1155/2022/6852022


https://doi.org/10.1016/j.crfs.2021.03.009


https://doi.org/10.1016/j.tifs.2016.10.014


https://doi.org/10.3390/foods12061242

https://doi.org/10.3389/fgene.2019.00524
https://doi.org/10.18637/jss.v033.i01
https://doi.org/10.1007/s10260-021-00579-1
35. Scott SL. BoomSpikeSlab: MCMC for Spike and Slab Regression. R package v. 1.2.5; 2022. Available from: https://CRAN.R-project.org/package=BoomSpikeSlab
Fig. 1. Systems view of causal AI model application for process and market decision making, management and innovations of food quality by do(x) inference

Fig. 2. Prediction of the wheat baking quality (volume V) with the 10 key features by the random forest model
P: protein (% DM)
WG: wet gluten (%) 
FN: falling number (s) 
V: bread loaf volume (cm³) 
WA: water absorption (%) 
R: dough resistance (min.) 
R/Ext: ratio resistance/extensibility 
TGT: total glutenin (%) 
THMW: total HMW (%) 
a: alpha gliadin (%) 
DS: dough degree of softening 

**Fig. 3.** Causal Bayes network model of the wheat key features and bread baking quality as volume V. The path coefficients are the direct causal strengths evaluated with the standardized variables 

**Fig. 4.** Average direct causal effects (ACE) of the wheat key features on bread baking quality. The ACE values are evaluated with the standardized variables
Fig. 5. Distributions of a bread loaf volume $V(\text{do}(x))$ caused by intervention $\text{do}(x)$ in content of THMW total high molecular gliadins (A) and P protein (B)

Fig. 6. DAG (directed acyclic graph) of causal effects milk composition and process parameters on consumer assessment of dairy quality
Please note that this is an unedited version of the manuscript that has been accepted for publication. This version will undergo copyediting and typesetting before its final form for publication. We are providing this version as a service to our readers. The published version will differ from this one as a result of linguistic and technical corrections and layout editing.

**Fig. 7.** Distributions of quality(do(x)) of consumer assessment of dairy quality caused by change in pre-treatment temperature (A) and fat content (B)

**Fig. 8.** Quality wine ACE caused by change of the standardized values of physical and chemical parameters