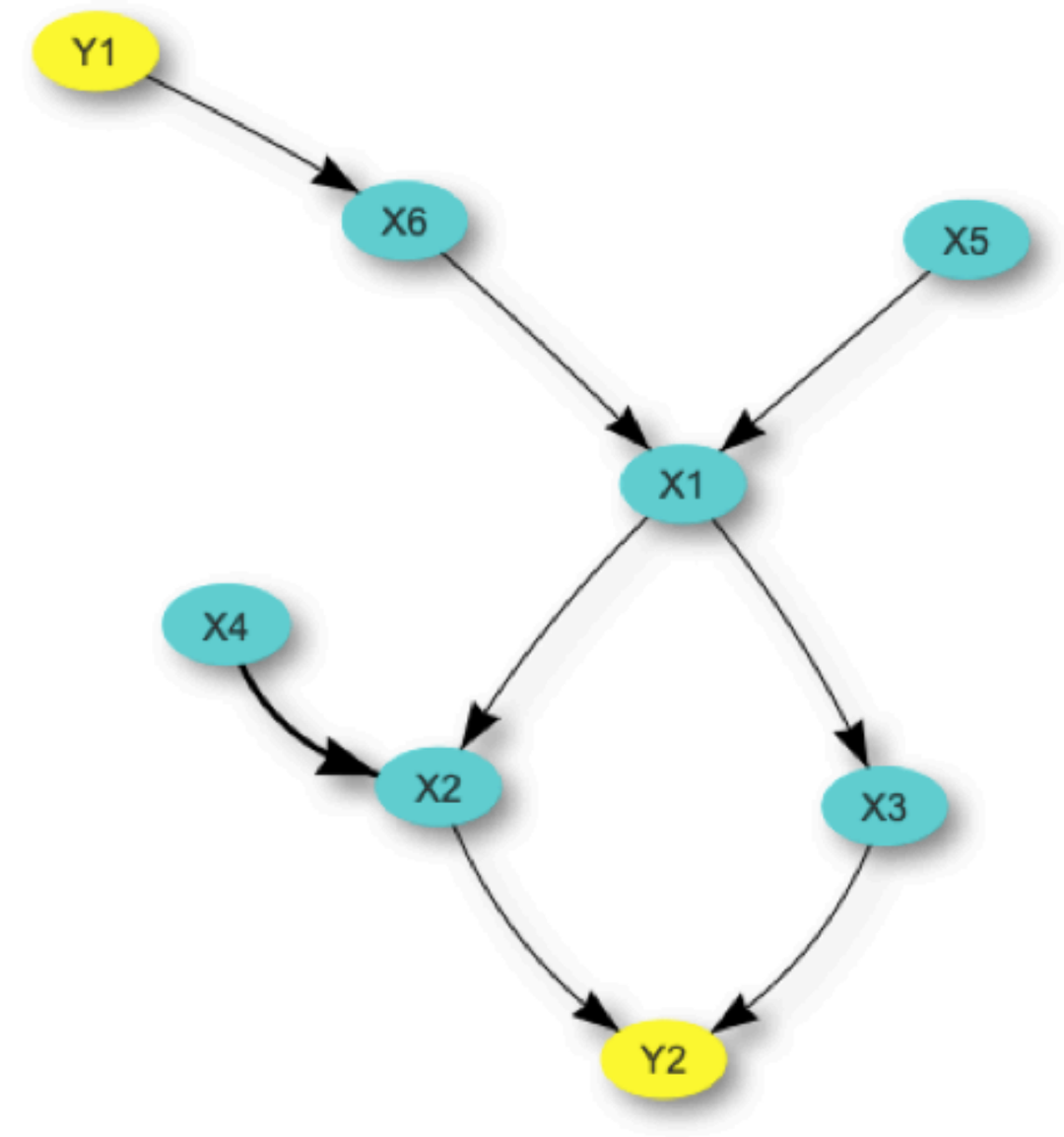
Bayesian Networks: elicitation, temporal modeling and applications

Thaís C. O. Fonseca - DME\UFRJ, Brazil
thais@im.ufrj.br
sites.google.com/site/thaisf

Joint work with Martine J. Barons (AS&RU, Department of Statistics, University of Warwick), Jim Q. Smith (AS&RU, Department of Statistics, University of Warwick), Hannah Merwood (Government Operational Research Service, UK), Alex Green (The National Archives, UK) and David H. Underdown (The National Archives, UK), J. T. Alves (epidemiologist, ANS Brazil), Maria do Carmo Leal (Epidemiologist, Fiocruz), Tatiana H. Leite (Epidemiologist, UERJ), Rosa Domingues (Epidemiologist, Fiocruz), Kelly Gonçalves (DME\UFRJ), Guilherme Oliveira (CEFET-MG) and Luiz Eduardo S. Gomes (PhD student, IM, UFRJ).

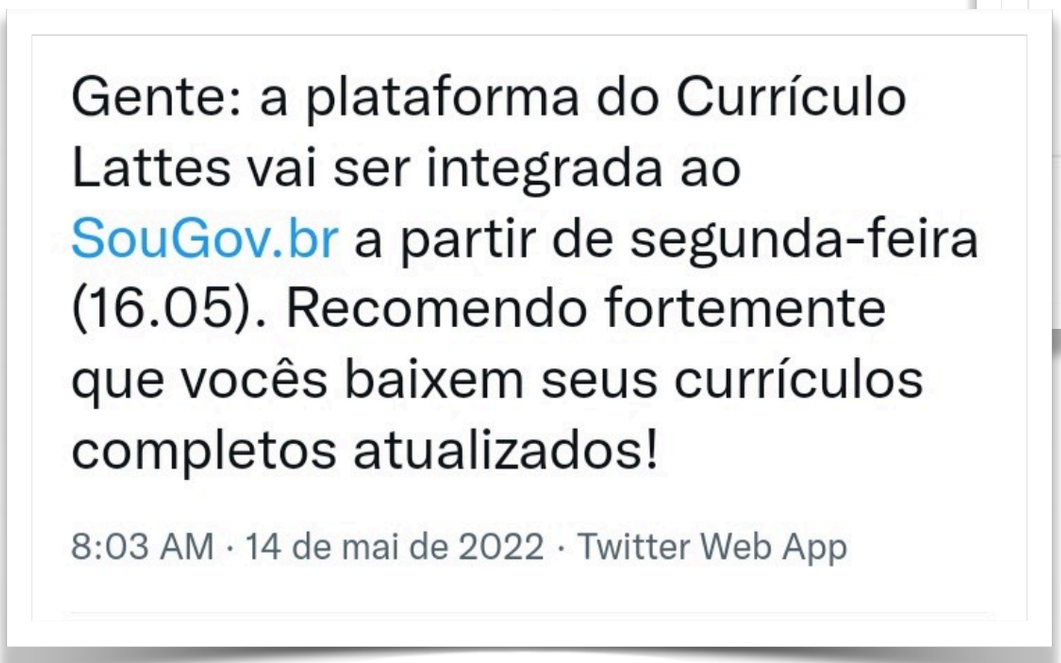
Outline

- Probabilistic Expert systems and Decision support based on BN
- Probabilistic graphical modeling: BN, divide and conquer
- Soft elicitation and elicitation of probabilities
- The digital preservation network construction and policy comparisons
- Some ongoing projects: food security, birth care system and robust graphical analysis



Can you render your Lattes?

- In 2021, news indicated that the Lattes webpage was unstable or not accessible.
- Is it safe to have only one copy of your Lattes?



Why is Digital Preservation complex?

- Digital records comprise primary sources which may be ***physical***, ***born digital*** or ***digitised***.

- They are fluid and fragile data: web-based tools, video, websites, structured datasets with multiple creators and owners.

- And it's under threat from

  - Rapidly evolving technology;

  - Outdated policies and standards;

  - A skills gap across the archives sector.

Digital records rely upon short-lived software and hardware for their survival and will rarely last even a decade without intervention.

Do you have enough security in your system? Can you ensure authenticity?

Expert systems

- A system that performs intellectually demanding tasks and that depends on an ability which is restricted to a particular area of expertise is called **expert system**.

- By formulating the expert's knowledge in an appropriate **formal (computer) language**, the reasoning conducted by the expert can be carried out by a computer.

- Here we are not interested in replacing the experts with AI, we want to build **a support decision tool** to aid the decision making in complex problems.

- In particular, we will consider **Probabilistic Network** as the formal language to construct our decision support system.

Uncertainty in Expert Systems

- Consider the medical expert system and cause-effect relations:

Smoking → Bronchitis → Dyspnoea

- These rule-based systems with **certainty factors** have serious limitations!

- Note that only a proportion of the smoking patients suffer from bronchitis. And dyspnoea appears as a symptom only for some of the patients with bronchitis.

- The majority of cause–effect mechanisms of interest in our attempts to model parts of the world in expert (or AI) systems are **uncertain**.

- Thus, we focus our attention on a method based on a **probabilistic interpretation of certainty factors**, leading to the definition of **Bayesian networks** (Kim & Pearl 1983, Pearl 1988).

Decision support

- In ever-larger systems, such as the digital preservation, it is increasingly difficult for decision makers to effectively account for all the variables within the system.

- It is well known that the human brain, when faced with too many alternatives, is not able to choose the optimal option.

$$\tilde{A} = arg\ max_{A \in \mathcal{A}} E[U(\tilde{\mathbf{y}}; A)]$$

- Tversky & Kahneman (1981) have shown that people usually do not make decisions that maximize their expected utility!

- Thus supporting human decisions by recommendations from ***decision support systems*** can improve the quality of decisions.

Back to the digital preservation system: What do you mean with: "*my file is preserved!*"

- *Preservation is meaningless without* **access** - Julian Morley (Stanford Libraries)
    - It is useless to have an intact record if you cannot access the file.
- What about a floppy disc? Is it preserved? It depends on you having the **tools** to render its content…
    - Interdependence between software, hardware, data and skilled archivists is increasingly complex.

Renderability and Intellectual control

- Digital preservation is measured on whether material is **renderable** and whether the archivist has full **intellectual control**.

- Some important questions regarding digital archives:

    - “If we improve our **system security**, how much will this decrease our risk?”

    - “Should we prioritise diversifying our **storage** media?”

    - Should we move our data/records to an **external storage**? Examples include Amazon Simple Storage Service, Microsoft Azure Archive Storage and Google Cloud Storage.

Main goals

- The online tool (**DiAGRAM**) aims to:
    - Improve users' understanding of the complex digital archiving risks and of the interplay between risk factors.
    - Empower archivists to compare and prioritize different types of threats to the digital archive: from software obsolescence to natural disaster.
    - Aid in quantifying the impact of risk events and risk management strategies on archival outcomes to support decision making.

Steps of our BN framework

(1) Construct the network structure (define all variables and connections) based on **soft elicitation**;

(2) Elicit the conditional probability tables when data is not available based on the **IDEA protocol**; or estimate parameters of conditional probability distributions.

(3) Estimate marginal and joint probabilities (**BN model**);

(4) Obtain the expected utility for each policy (**Logic sampling**);

(5) Compare utilities which will be available to aid the decision maker.

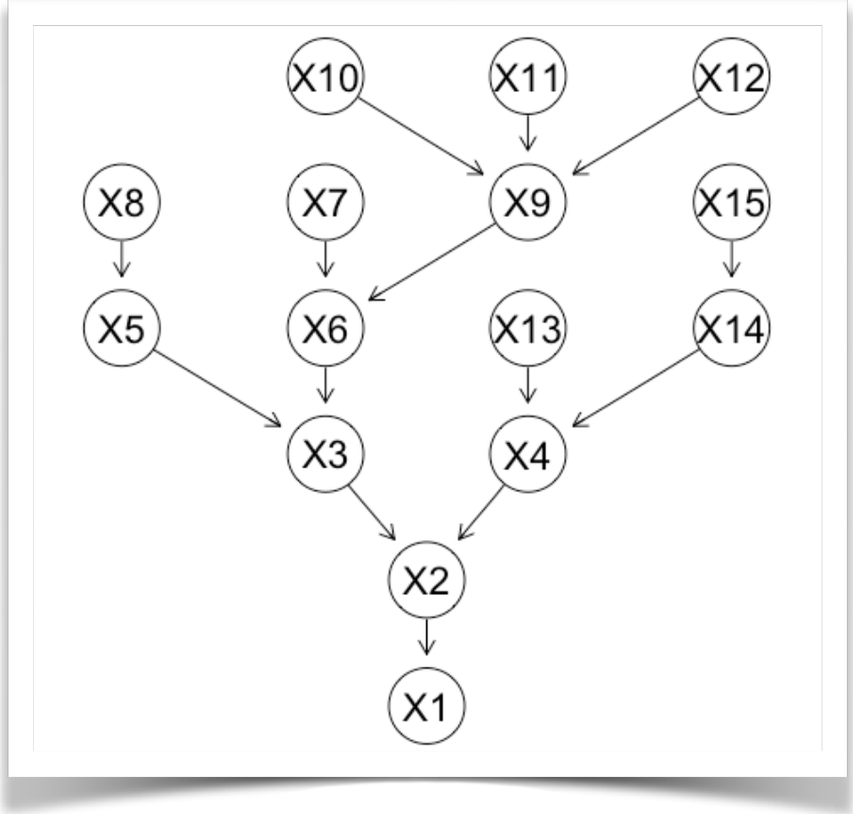
DiAGRAM

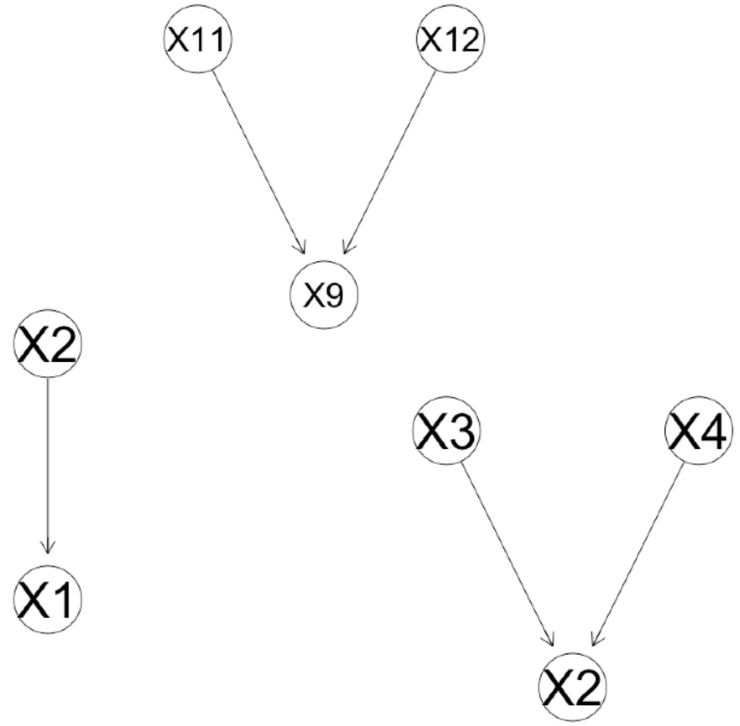
Bayesian Networks

Bayesian network: divide to conquer

- BN are **graphical models** that can represent causal relations among variables.

- We consider the idea of conditional probabilities to divide a large multivariate problem in smaller ones based on **conditional independence**.





Example: Lung diseases (Lauritzen and Spiegelhalter, 1988)

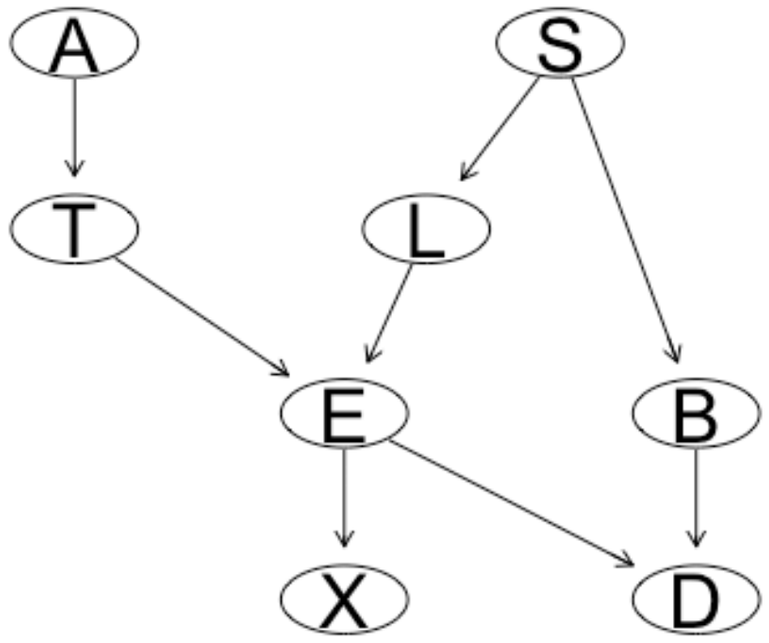
- Divide to conquer: we can use **parallel computing** for large tasks and we have much less parameters.

- Example: data set from Lauritzen and Spiegelhalter (1988) about lung diseases and visits to Asia.

A: visit to Asia? S: smoking?

T: tuberculosis? L: lung cancer? B=bronquitis?

E: either T or L? X: positive X-ray? D: dyspnoea?



- Parameters in the joint model: 2^8-1 = 255.
- Parameters in the BN model: 18

Graphical models

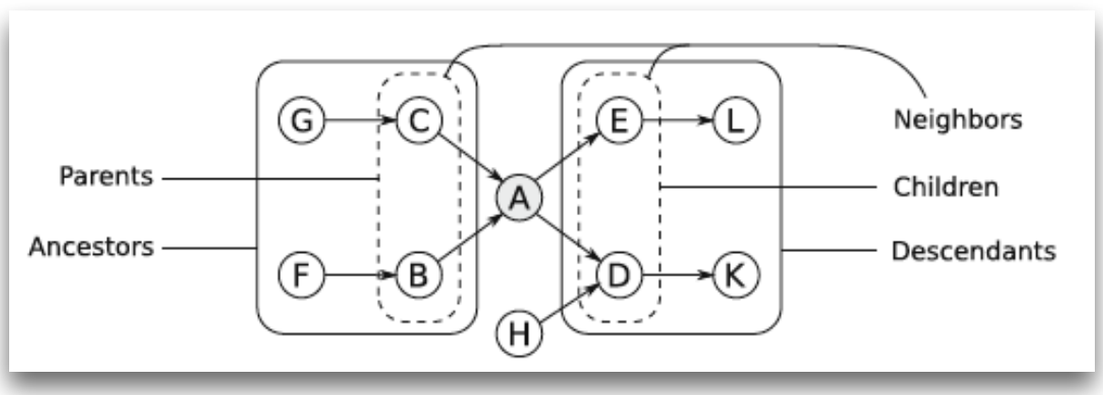
A probabilistic network model is a directed acyclic graph (DAG) with the vertices representing variables and the edges representing relations among the variables.

A graph $\mathcal{G}$ is a mathematical object with

- nodes $V = \{v_1, \ldots, v_p\}$;
- arcs A, so that $a_{ij} = (v_i, v_j)$.

A Bayesian network model will assume vertices $\mathbf{Y}$ such that

$$\mathbf{Y}_i \perp \mathbf{Y}_j \mid \mathbf{Y}_{\Pi_i},\ i, j = 1, \ldots, p,\ j \notin \Pi_i.$$



D is condicionally independent of E given A.

Bayesian network (BN)

- A Bayesian network (Pearl, 1988) expresses human-oriented qualitative structure translated into a joint probability distribution for the vector **Y** = (Y_1,…,Y_p).

$$f(y \mid \mathcal{G}, \theta) = \prod_{i=1}^{p} f_i(y_i \mid Y_{\Pi_i}, \mathcal{G}, \theta_i),$$

with Y_{Π_i} the parents of y_i.

Local Markov property

- A BN is defined by two basic elements:

A set of conditional independence statements, represented by the GRAPH.

+

A set of local conditional distributions.

Learning Bayesian Networks

Learning Bayesian Networks

Model selection and estimation of BNs are known as *learning*, and are usually performed as a two-step process:

- structure learning, estimating the graph from the data (or expert knowledge);

- parameter learning, estimating the local distribution parameters given the graph learned in the previous step from the data (or expert knowledge).

Structure Learning

Many conditional independence tests are performed. Computational complexity is super-exponential in the number of nodes (p) in the worst case.

- **Structure learning:** identifying the graph of the BN. It should be the minimal map of the dependence structure of the data.

- **Causal inference**: Learning the structure is useful, because we can use structure to infer causal relationships, and consequently predict the effects of interventions in the outcome of interest.

- **Possible routes**: constraint-based, score-based, hybrid algorithms or **elicitation based on expert knowledge**.

Parameter learning

- The use of local conditional distributions alleviates the curse of dimensionality.
- The three most commom choices for local distributions are
  - Discrete BN: $Y_i \mid Y_{\Pi_i}$ is Multinomial;
  - Continuous BN: $Y_i \mid Y_{\Pi_i}$ is Gaussian;
  - Hybrid BN: $Y_i \mid Y_{\Pi_i}$ is a mixture of Gaussian distributions for each level of a discrete valued parent;

Discrete BN

- We assume the multinomial model such that

$$\mathbf{Y}_i \mid \mathbf{Y}_{\Pi_i} = \mathbf{z}_j, \boldsymbol{\theta}_{ij} \sim Mult(M_{ij}, \boldsymbol{\theta}_{ij}),$$

  with M_{ij} the counts of $\{\mathbf{Y}_{\Pi_i} = \mathbf{z}_j\}$, θ_{ijk} the probability that $\mathbf{Y}_i$ is in state k given that the parent set is in state j, $\theta_{ijk} > 0$ and $\sum_{k=1}^{n_i} \theta_{ijk} = 1$.

- If a Dirichlet prior with parameter $\mathbf{a}_{ij}$ is assumed for $\boldsymbol{\theta}_{ij}$ then the posterior distribution is Dirichlet with density

$$\rho(\boldsymbol{\theta}_{ij} \mid N_{ij}, \mathbf{z}_j, B) = c_i \prod_{k=1}^{n_i} \theta_{ijk}^{N_{ijk}+a_{ijk}-1},$$

  with N_{ijk} the counts of $\{Y_{ik} = y_{ik}\}$ when $\{Y_{\Pi_i} = z_j\}$, $i = 1, \ldots, n_i$, $j = 1, \ldots, q_i$, and $M_{ij} = \sum_{k=1}^{n_i} N_{ijk}$.

Posterior inference and prediction (Heckerman et al., 1995)

Global independence: parameters associated with each variable in the network are independent;

$$\Theta = \cup_{i=1}^{p} \Theta_i$$

Local independence: parameters associated with each state of the parents of a variable are independent;

$$\Theta_i = \cup_{j=1}^{q_i} \Theta_{ij}$$

These two assumptions together make **computation fast and scalable** to large networks.

D. Heckerman et al., Learning Bayesian Networks: The Combination of Knowledge and Statistical Data, Machine learning 20, 197-243, 1995.

Approximate inference: Logic sampling

- Utility of competing policies are compared computing the predictive distribution which is approximated using the logic sampling (Monte Carlo simulation).

- ***Basic idea***: to sample from a BN we transverse the network in topological order, visiting parents before children and generate a value of each visited node according to the conditional probability of that node.

$1-$ Order the variables in topological order $Y_{(1)} \prec \ldots \prec Y_{(p)}$;

$2-$ Set $n_{\mathcal{E}}=0$ and $n_{\mathcal{E},\mathcal{Q}}=0$;

$3-$ Repeat for $m = 1, \ldots, M$:

$3.1-$ Generate $Y_{(i)}$ from $Y_{(i)} \mid Y_{\Pi_{(i)}}$;

$3.2-$ If $\mathbf{y}$ includes $\mathcal{E}$, set $n_{\mathcal{E}} = n_{\mathcal{E}} + 1$;

$3.3-$ If $\mathbf{y}$ includes $\mathcal{E}$ and $\mathcal{Q}$, set $n_{\mathcal{E},\mathcal{Q}} = n_{\mathcal{E},\mathcal{Q}} + 1$;

$4-$ Estimate $p(\mathbf{y}_q = \mathcal{Q} \mid \mathcal{E}, \mathcal{G}) = n_{\mathcal{E},\mathcal{Q}}/n_{\mathcal{E}}$.

Structure learning (Soft elicitation)

Soft elicitation

- The *facilitated modelling*: analysts, problem owners and experts meet in workshops to ‘solve’ the problem.
    - What are the processes, inputs, outputs, actors etc
    - How do these entities interact?
    - What are the uncertainties?
    - How might these be modelled?
    - What relevant data and expertise are available?

See: S. French (2021) From soft to hard elicitation. Journal of the Operational Research Society, 1–17.

Expert panel in elicitation workshop and meetings

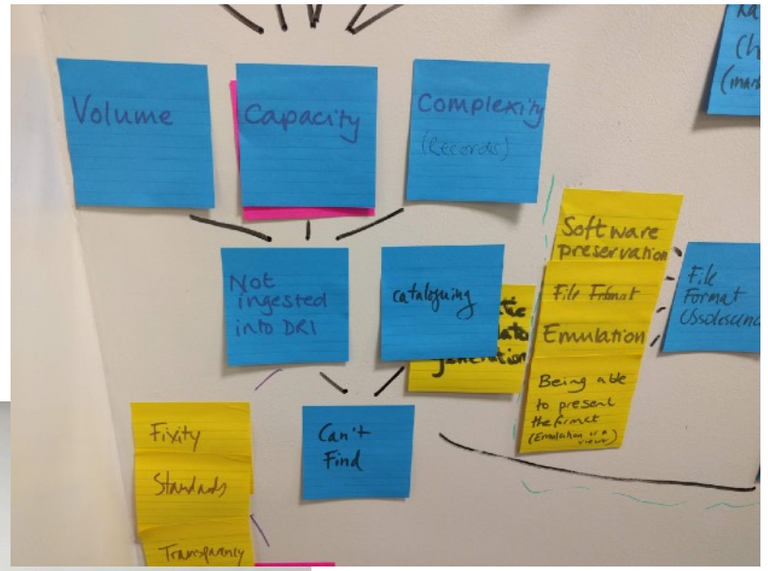
What is an ***expert***? According to O'Hagan et al 2006, it is someone who has great knowledge of the subject matter.

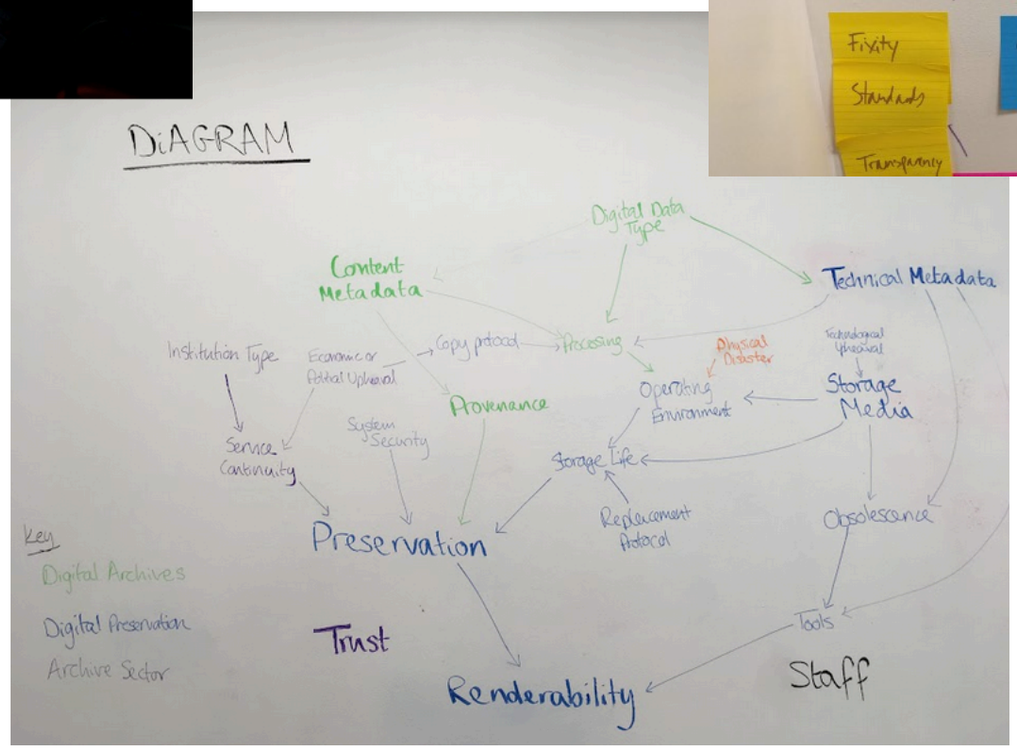
Partners in elicitation workshops: Delegates from The National Archives; The Applied Statistics & Risk Unit, University of Warwick; Dorset History Centre; Gloucestershire Archives; TfL Corporate Archives; Special Collections (University of Leeds); Design Archives (University of Brighton); The Digital Preservation Coalition.

These delegates were engaged in what can be called joint model building or ***soft elicitation*** (Wilkerson, 2021; French, 2021) and also in the ***probability elicitation*** workshops.

Expert panel in elicitation workshop and meetings

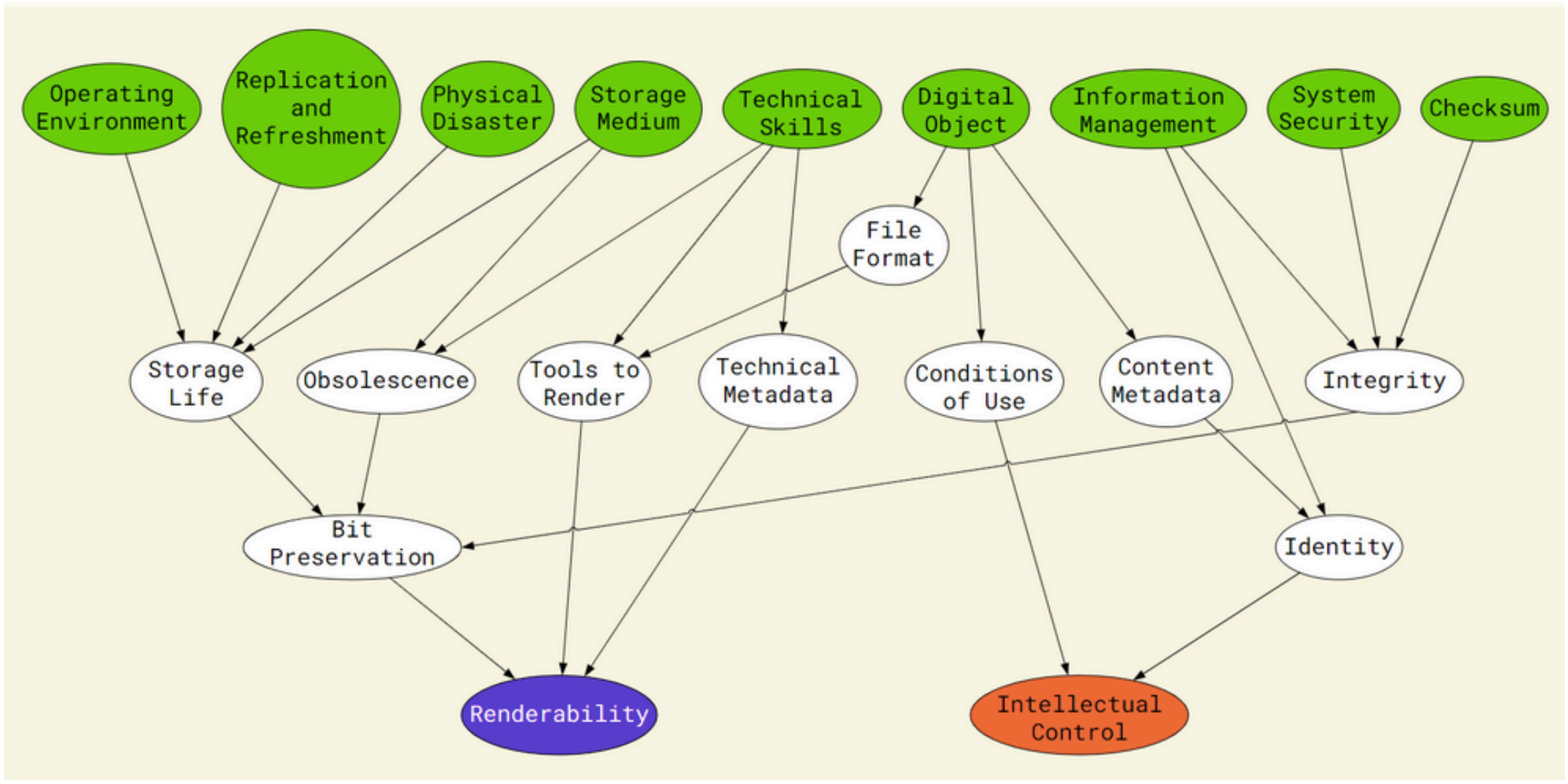
- A series of workshops was held which iteratively and collaboratively produced a consensus on the interrelations between various subsystems and data that was available.





The network has **19 + 2** variables

9 variables are customizable to reflect each institution



Renderability and **Intellectual control** are the nodes of interest used to compute utility of policies.

Digital preservation network

Qualitative system representation

Elicitation of probabilities (IDEA protocol)

Why expert elicitation?

- Often decision need to be made in situations of rare events or sparse data.

- These are the situations in which decisions become particularly difficult to be made.

- Our solution to this problem is to **synthesise the opinions of a group of experts**.

- Methods for elicitation of beliefs under uncertainty: **SHELF** (Oakley and O'Hagan, 2016) e **IDEA²** (Hanea et al., 2017).

[2] Hanea, A., McBride, M., Burgman, M., Wintle, B., Fidler, F., Flander, L., Twardy, C., Manning, B., Mascaro, S., 2017. Investigate discuss estimate aggregate for structured expert judgement. *Int. J. Forecast.* 33 (1), 267–279.

IDEA protocol

- The protocol aims to motivate expert discussion, revision of opinions and to reduce biases in the elicitation of probabilities.

- Opinions of a group of experts can be combined by (1) group consensus or (2) mathematical aggregation. Here we consider aggregation.

Investigate

↓

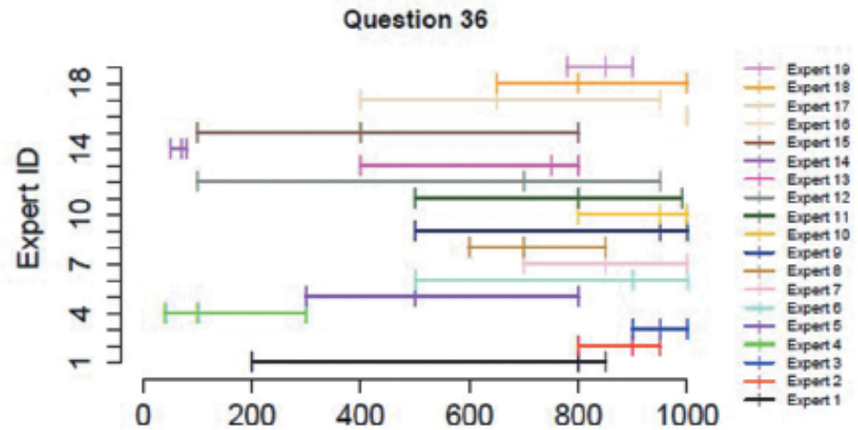
Discuss

↓

Estimate

↓

Aggregate



Pooling expert knowledge

- The Cooke's approach (Cooke, 1991) was used to pool the judgements of several experts into one commom aggregated probability distribution for each variable.

- The best performance experts will have larger weights in the pooled distribution.

- We defined **20** calibration questions and **24** target questions.

    - The calibration questions are questions related to experts' field, for which the true values are known to the statistician but not to the experts. It will access how accurate and informative each expert is.

- The experts are unaware which are the seed variable and the target questions.

Pooling expert knowledge

- Elicitation requires experts to make probabilistic judgements, such as median and quartiles, that are difficult and unfamiliar tasks for most experts. Training was given before the elicitation workshops.

- 22 experts participated in the elicitation workshops.

- Each expert provided quantiles (0.05,0.5,0.95) for each question.

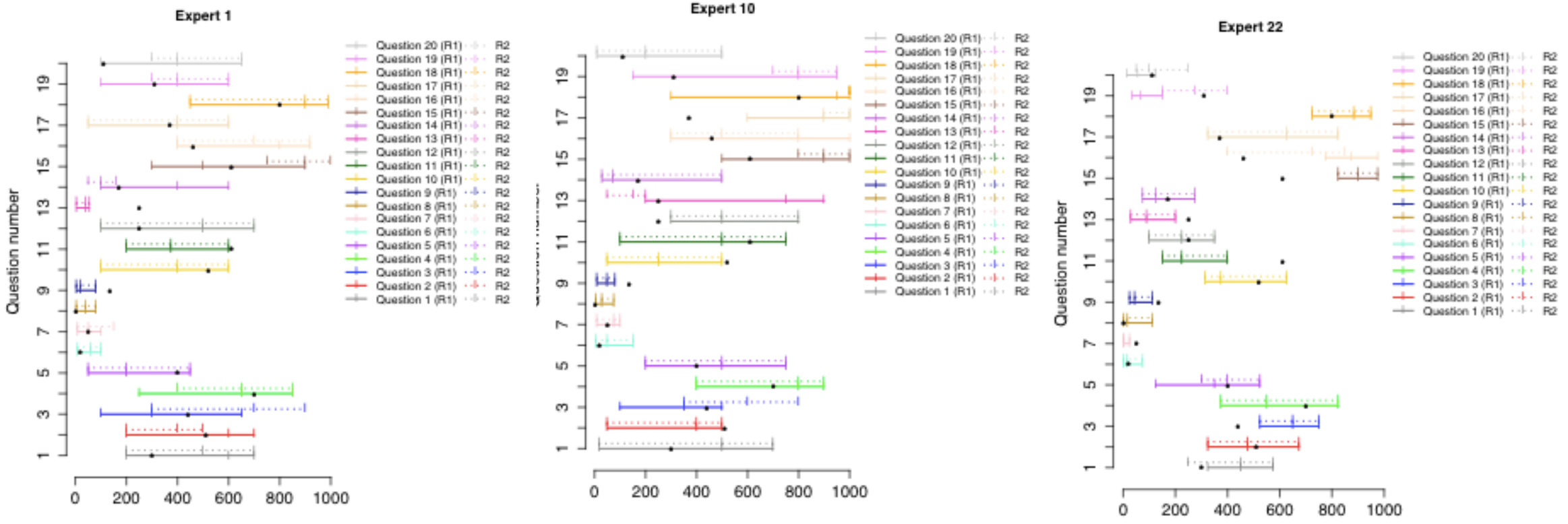
Calibration questions

- With the quantiles provided by experts, the proportion of times the true value was observed in each interval $(-\infty, q_{5\%})$, $(q_{5\%}, q_{50\%})$, $(q_{50\%}, q_{95\%})$, $(q_{95\%}, \infty)$ is counted.
- Note that the expected values for accurate and informative experts are $(5\%, 45\%, 45\%, 5\%)$.
- To measure the divergence between the distribution provided by experts and the theoretical model we consider the Kullback-Leibler (KL) divergence

$$KL(\mathbf{o}, \mathbf{p}) = \sum_{i=1}^{n} o_i \; log\left(\frac{o_i}{p_i}\right),$$

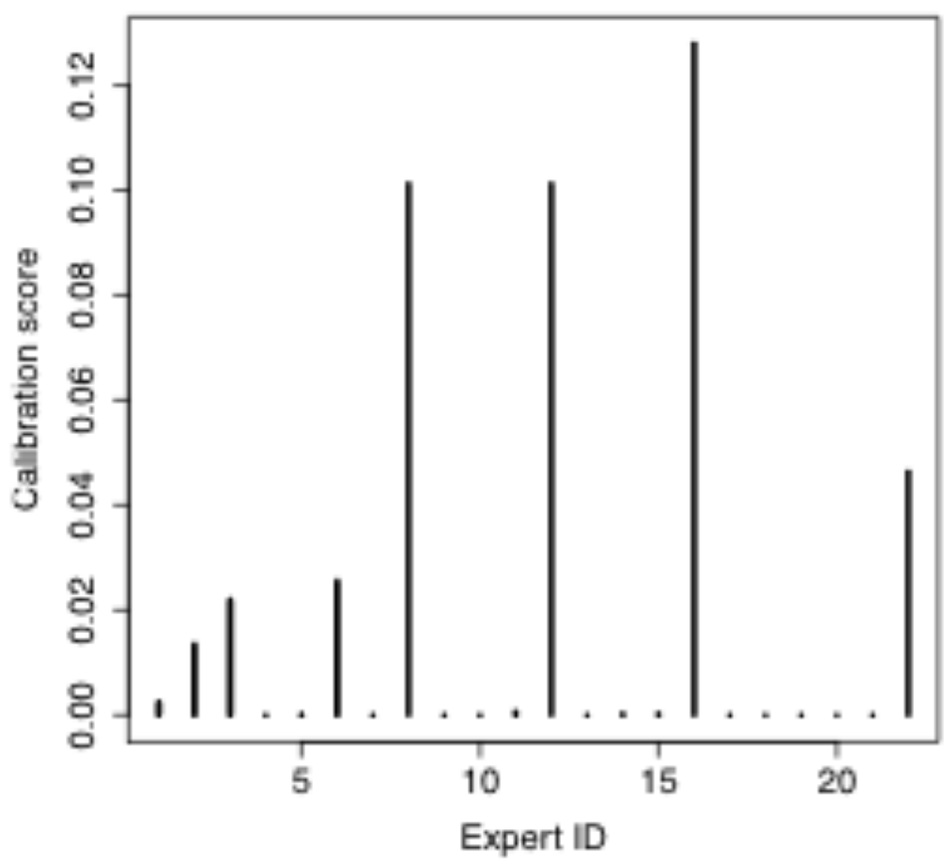
with o_i the observed proportion in interval i, p_i the theoretical probability of interval i and n the number of intervals.

Expert results for the calibration questions



Calibration score

- Based on the asymptotic distribution of KL we compute the calibration score for each expert, given by $P(2qKL(\mathbf{o}, \mathbf{p}) \geq k_{obs})$.
- If the expert is well calibrated then k_{obs} is small and the calibration score would be large.



Information score

- Issue: large intervals can lead to large calibration score but poor information.
- The spread of the experts' intervals is assessed relative to a reference interval (l_j, u_j), which is based on the range for all experts for question j.
- The KL divergence relative to $U_j \sim Unif(l_j, u_j)$ in the reference interval is computed. For the experts' intervals $\mathbf{I}_j = (l_j, q_{5\%,j}, q_{50\%,j}, q_{95\%,j}, u_j)$ we have that

$$P(I_{ji} < U_j < I_{j+1,i}) = \frac{I_{j,i+1} - I_{ji}}{u_j - l_j} = \tilde{o}_{ji}.$$

- The final information weight for each expert is

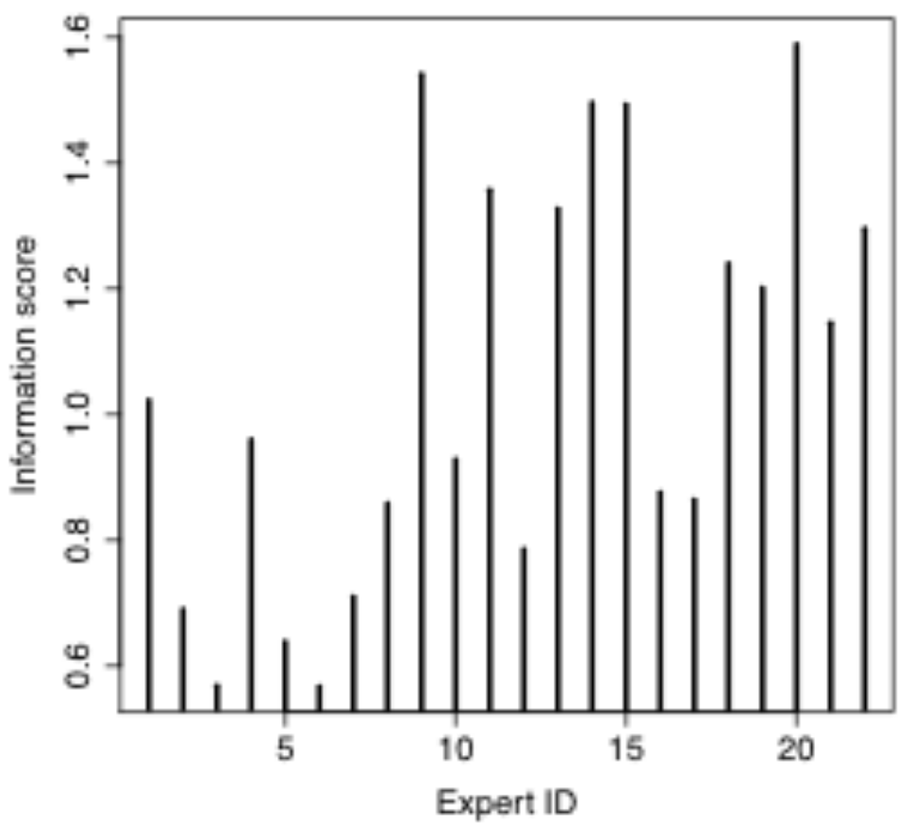
$$I = \frac{1}{q} \sum_{j=1}^{q} I(\mathbf{r}, \tilde{o}_j), \; I(\mathbf{r}, \tilde{o}_j) = \sum_{i=1}^{n} r_i log\left(\frac{r_i}{\tilde{o}_{ji}}\right),$$

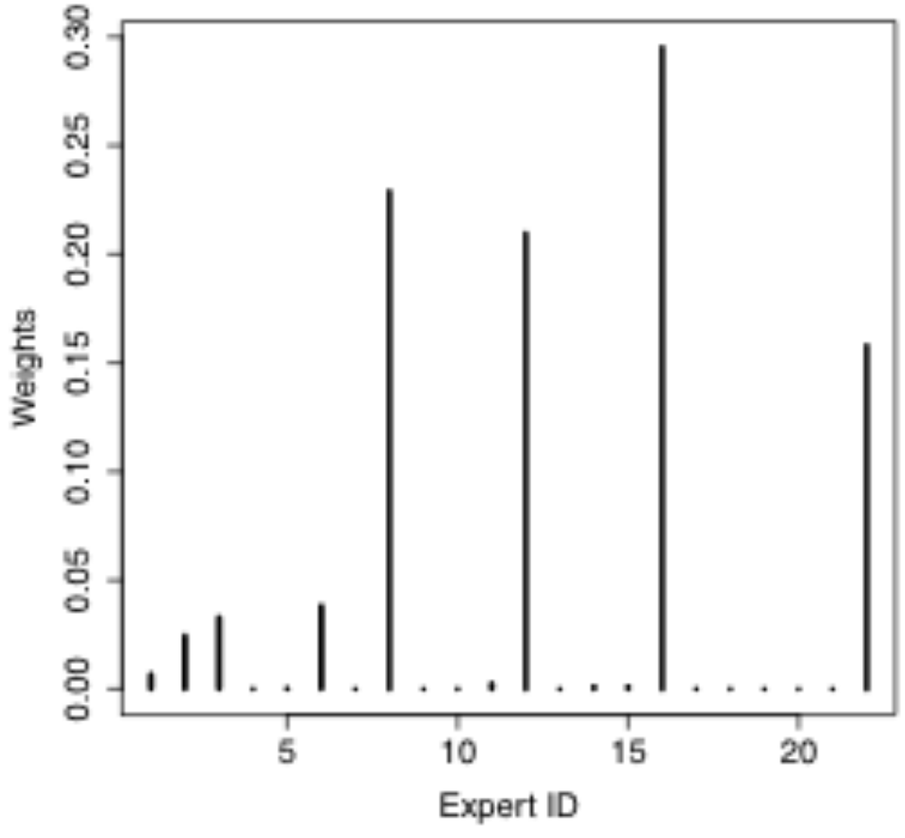
with $\mathbf{r}_j = (0.05, 0.45, 0.45, 0.05)$ the theoretical probability.

Information score and final weights

- The weight is computed as the product of the calibration and the information scores.

- Scores close to 0 indicate worse performances.





Aggregate

- Here we consider mathematical aggregation. The combined distribution is obtained as

$$P(X \leq x) = \sum_{i=1}^{n_e} w_i P_i(X \leq x),$$

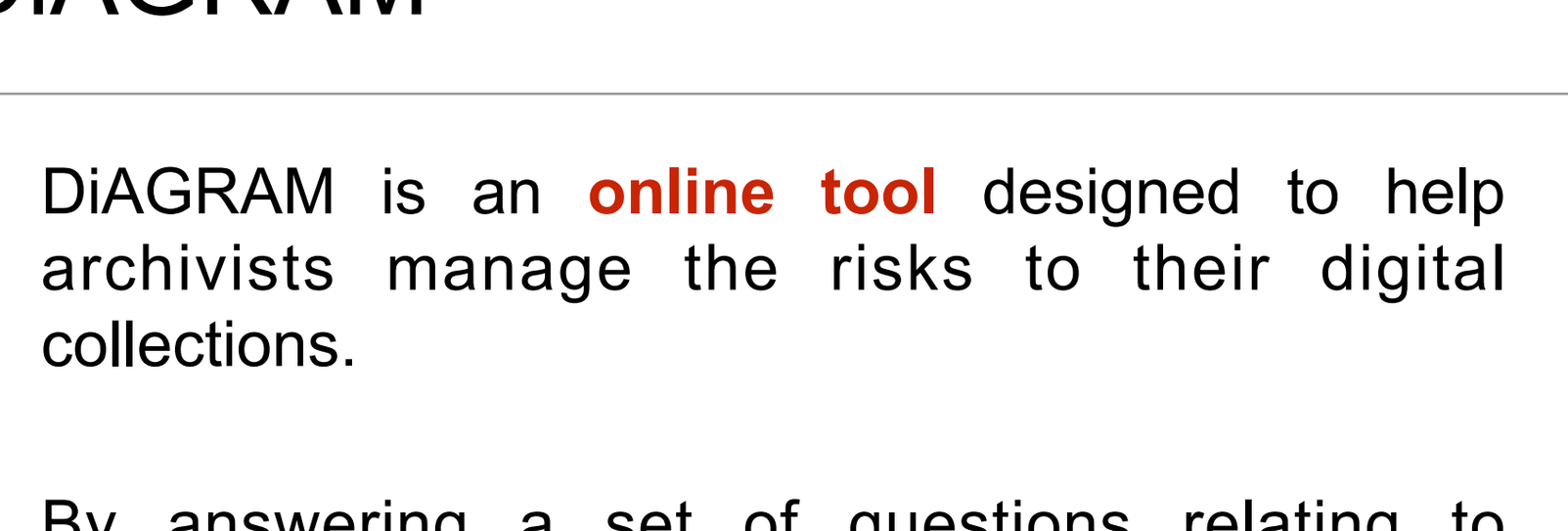
$P_i(X \leq x)$ is the cumulative distribution provides by expert i, n_e is the number of experts.

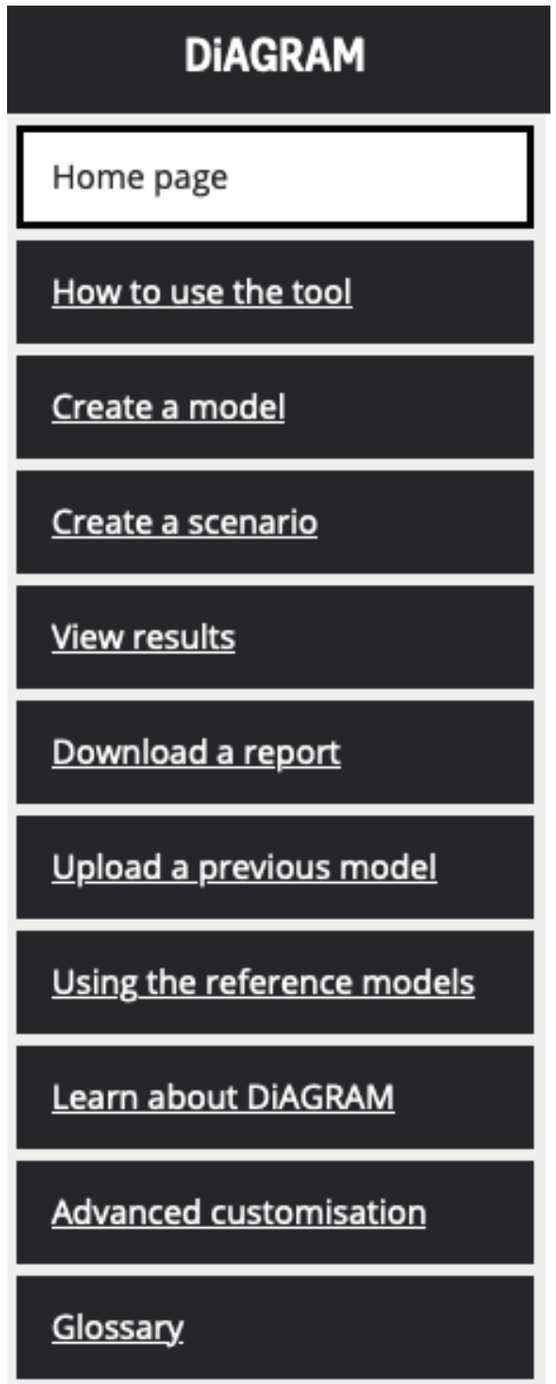
DiAGRAM and Policy comparison

DiAGRAM

- DiAGRAM is an **online tool** designed to help archivists manage the risks to their digital collections.

- By answering a set of questions relating to archives, the tool will calculate the probability that your digital material is preserved, compare scenarios and policies.

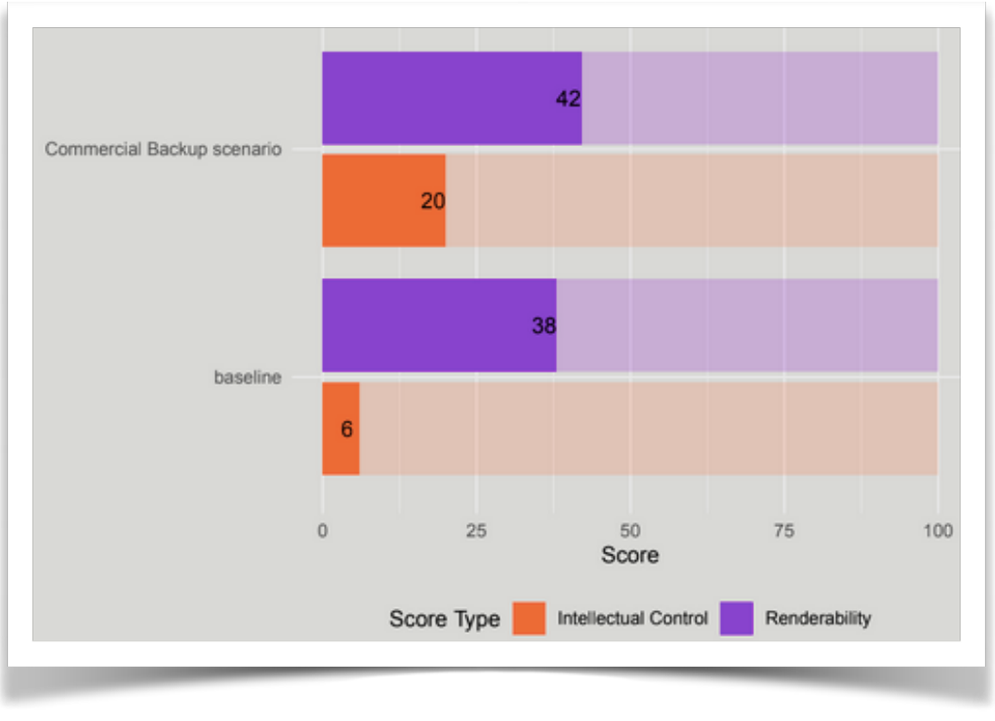




Utility comparison

- For comparative purposes DiAGRAM provides a Baseline Model (BM).

- We compare the BM with the alternative scenario of Commercial Backup (CB), which is as BM but improving information management to 43% and technical skill level to 30%.



Resources

- Safeguarding the Nation's Digital Memory: Towards a Bayesian Model of Digital Preservation Risk (**2021**) M J Barons, S Bhatia, T C O Fonseca, A Green, S Krol, H Merwood, A Mulinder, S Ranade, J Q Smith, T Thornhill, D H Underdown *Archives and Records*, 42:1, 58-78.

- Project webpage: www.nationalarchives.gov.uk/information-management/manage-information/preserving-digital-records/research-collaboration/safeguarding-the-nations-digital-memory/

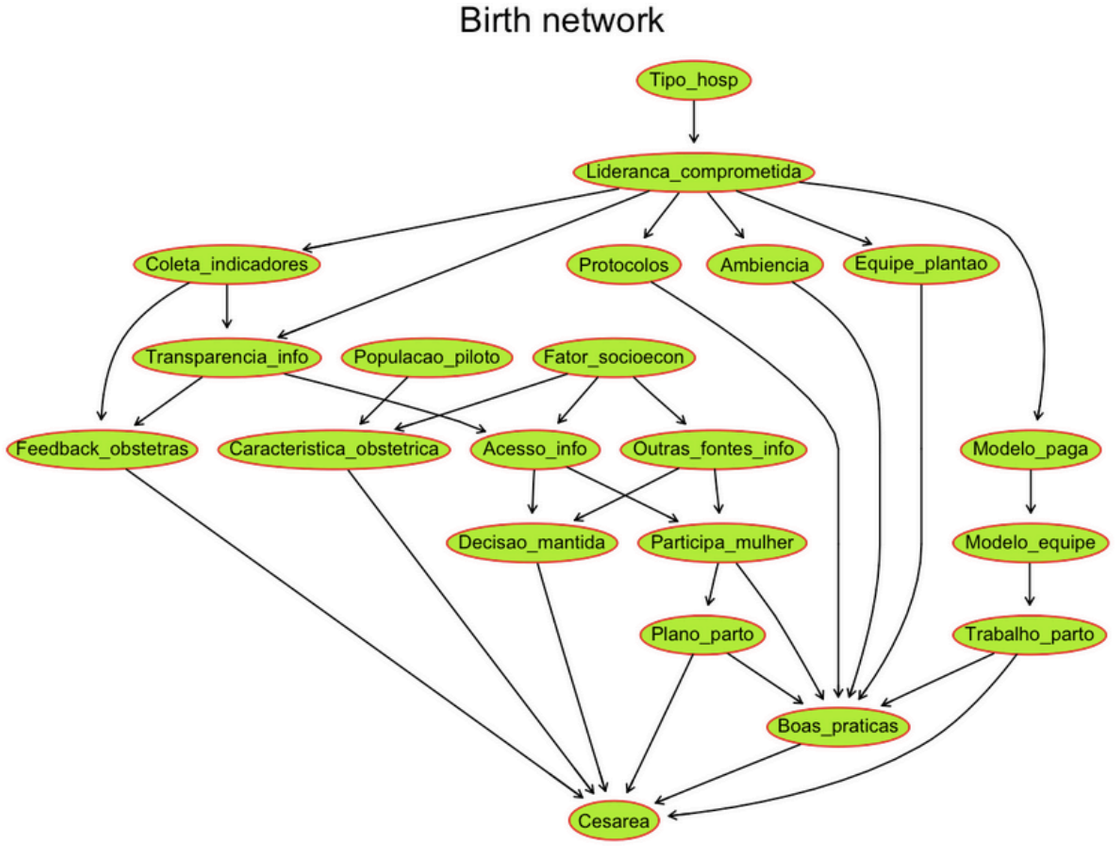
- **DiAGRAM app**: https://nationalarchives.shinyapps.io/DiAGRAM/

- Funding: This work was supported by the National Lottery Heritage Fund under project reference number OM-19-01060; The Engineering and Physical Sciences Research Council under grant EP/R511808/1;

Other applications of Bayesian Networks

Implementation analysis of Adequate Childbirth Project (PPA): a Bayesian network strategy.

Joint work with: J. T. Alves (epidemiologist, ANS Brazil), Maria do Carmo Leal (Epidemiologist, Fiocruz), Tatiana H. Leite (Epidemiologist, UERJ), Rosa Domingues (Epidemiologist, Fiocruz)



- Brazil is one of the countries with the highest prevalence of cesarean-sections in the world. Rates up to **88%** in the private hospitals.

- Evidence indicates that cesarean rates above **15%** are related to maternal mortality, mobility, "near death" of babies, metabolic syndromes, diabetes and asthma.

- A quality improvement project called “Projeto Parto Adequado”(PPA) was developed aiming to identify models of care for labor and childbirth.

- **Goals**: This research aims to evaluate the implementation of PPA in privates hospitals in Brazil using Bayesian Networks.

The Brazilian childbirth care system

- Several factors affect the outcome of interest and these factors are interrelated. A regression model would not be adequate in this setup.

- Data available: **4.289** women and more than **100** variables for each woman.

- Network construction: graph elicitation and predictive analysis (queries) to identify the most relevant factors. Team participating in the elicitation: obstetricians, nurses, mothers, epidemiologists and statisticians.

- Funder: EPSRC-funded GCRF Accelerator Account Fund (UK), from 01/12/2019 to 31/07/2020. UN Sustainable Development Goal: Good Health & Well-Being.

Food security modeling: DBN

Joint work with: Martine J. Barons (AS&RU, Warwick University), Andy Davis (Coventry & Warwickshire Local Enterprise Partnership), Jim Q. Smith (AS&RU, Warwick University).

- Food security exists when all people, at all times, have physical and economic access to sufficient, safe and nutritious food to meet their dietary needs and food preferences for an active and healthy life (FAO, 1996).

- Rising food insecurity has been associated with malnutrition, sustained deterioration of mental health, inability to manage chronic disease, worse child health (Loopstra et al., 2015a; Loopstra,2014) and it has been found to affect school children's academic performance, weight gain, and social skills (Faught et al., 2017).

- In this study, we consider as the main outputs of interest **malnutrition** and **school performance of children receiving free meals**.

- Goal: provide decision support for household food security in the UK.

Parameter learning with temporal dynamics

- In DBNs the time slices are connected through temporal links to form the full model which accommodates dependencies within and between time slices.
- Consider the general setting such that

$$\mathbf{Y}_{it} \perp \mathbf{Y}_{Q_i}^t \mid \mathbf{Y}_{\Pi_i}^t, \mathbf{Y}_i^{t-1},\ i = 1, \ldots, n,$$

with $\{\mathbf{Y}_t : \ t = 1, \ldots, T\}$ a multivariate time series composing a DAG whose vertices are univariate processes and Π_i the index parent set of Y_{it} and $\mathbf{Y}_i^t = (Y_{i1}, \ldots, Y_{it})'$ the historical data.

Each variable at time t depends on its own past series, the past series of its parents and the value of its parents at time t.

Parameter learning with temporal dynamics

- The observation and system equations are defined as a Multiregression Dynamical Model (Queen and Smith, 1993) and is given by

$$\begin{aligned} Y_{it} &= F_{it}\theta_{it} + \epsilon_{it}, \\ \theta_{it} &= G_{it}\theta_{i,t-1} + \omega_{it}, \end{aligned}$$

  with $\epsilon_{it} \sim N[0, V_{it}]$ and $\omega_{it} \sim N[0, W_{it}]$.

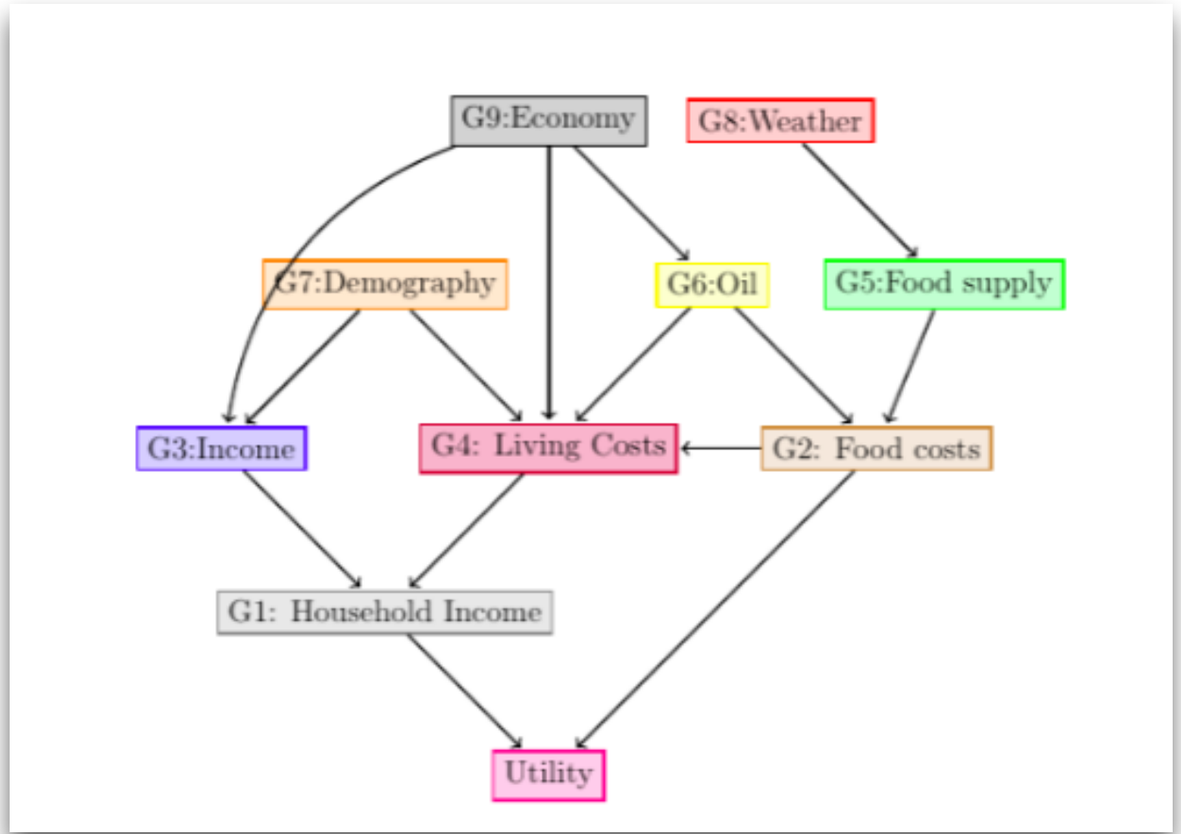
- Define $V_{it} = \phi_{it}^{-1}$, the variance evolution follows the gamma model given by

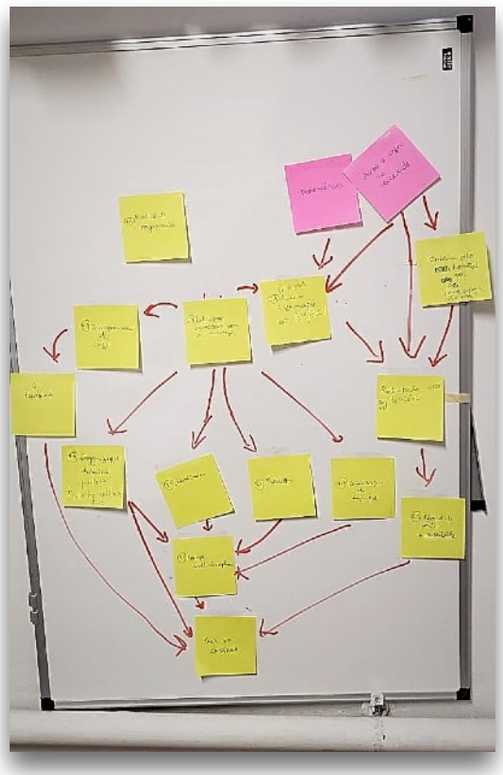
$$\phi_{it} \mid D_{t-1} \sim Gamma(\delta_i^* n_{i,t-1}/2, \delta_i^* d_{i,t-1}/2),$$

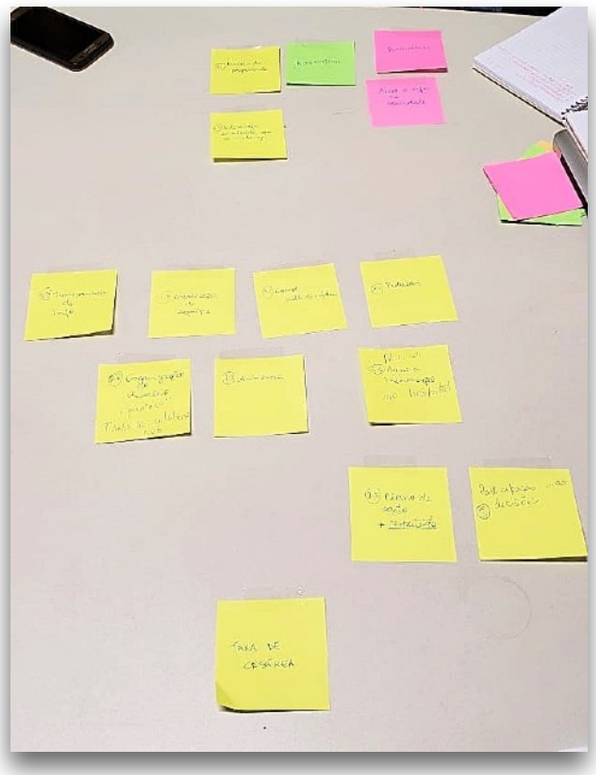
  with $\delta_i^* \in (0, 1)$ being the discount factors.

More details see: A decision support system for addressing food security in the UK. (**2021**) M.J.Barons, T.C.O.Fonseca, A.Davis and J.Q.Smith. Journal of the Royal Statistical Society, Series A (Statistics in Society), 00, 1– 24.

Expert panels

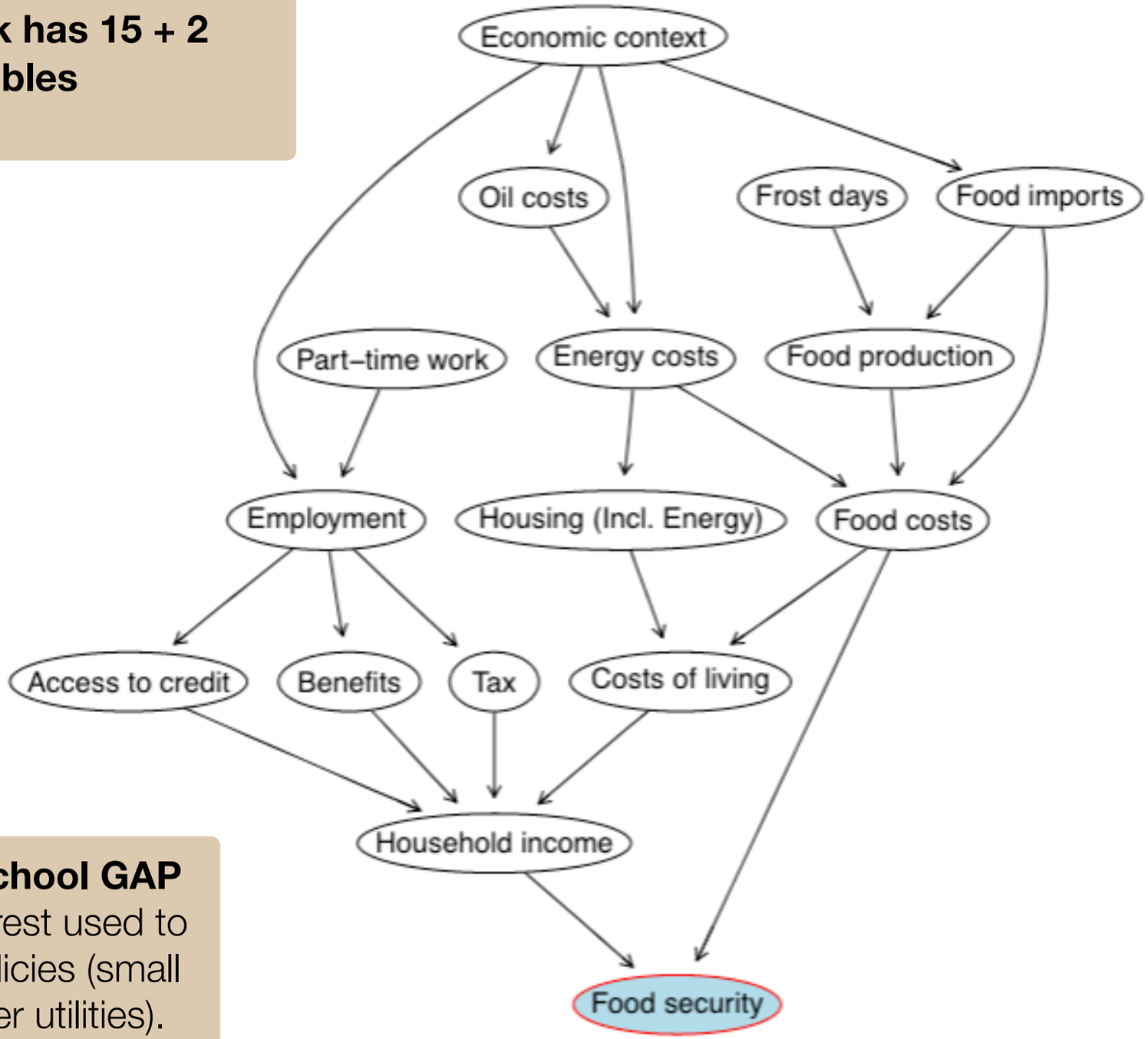






Experts in elicitation workshop: Delegates from the Warwickshire Council (public health, legal & governance, data and statistics, renewable energy, social & financial inclusion, localities & partnerships, child poverty, education, emergency planning, libraries & customer services, and corporate policy departments).

The network has 15 + 2 variables



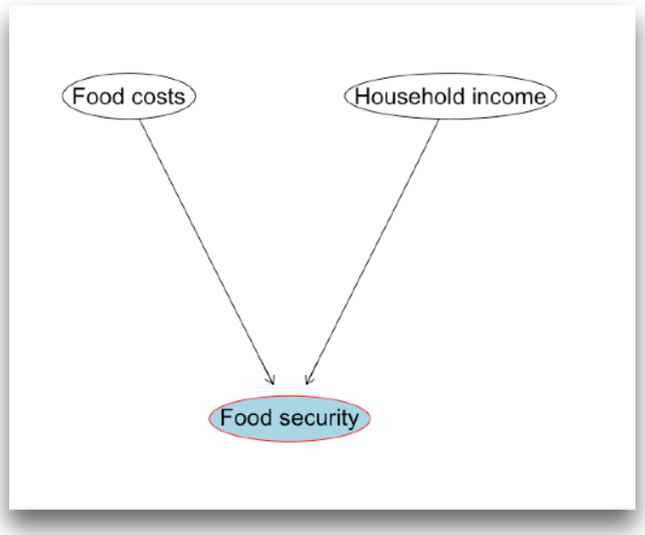
Malnutrition and **School GAP** are the nodes of interest used to compute utility of policies (small values result in better utilities).

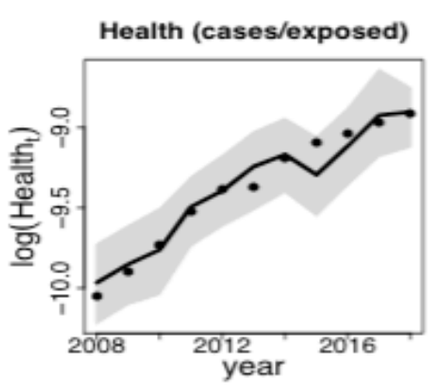
Food security network

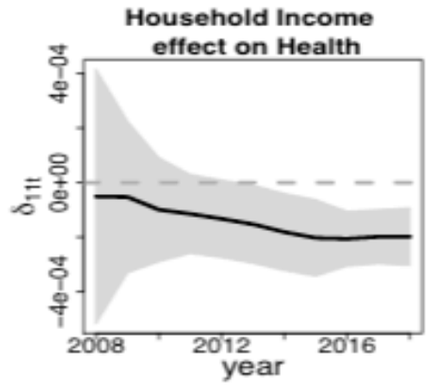
Qualitative system representation

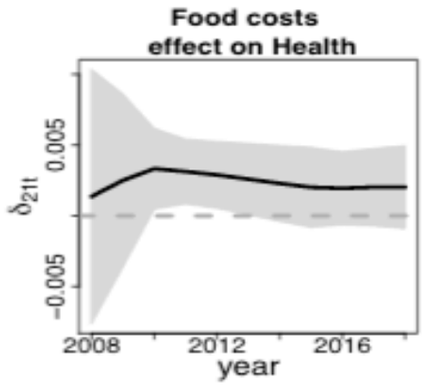
Sub-network analysis

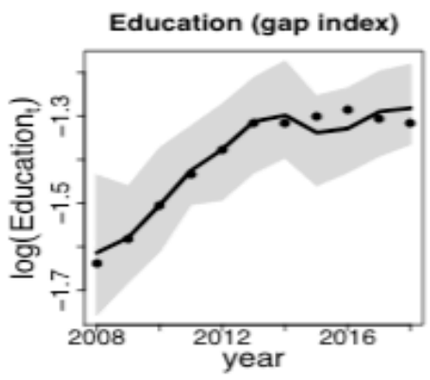
Consider the Food security variable and its parents (Household income and Food costs).

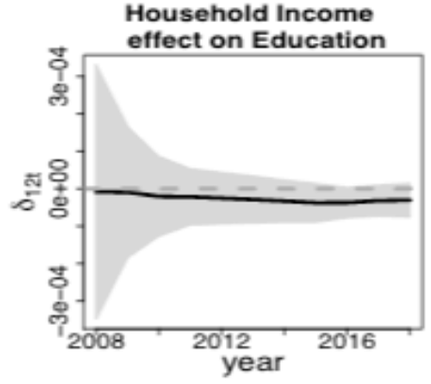


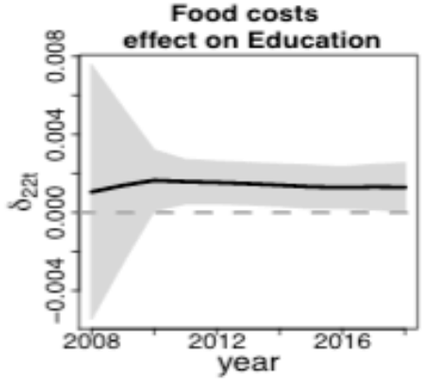












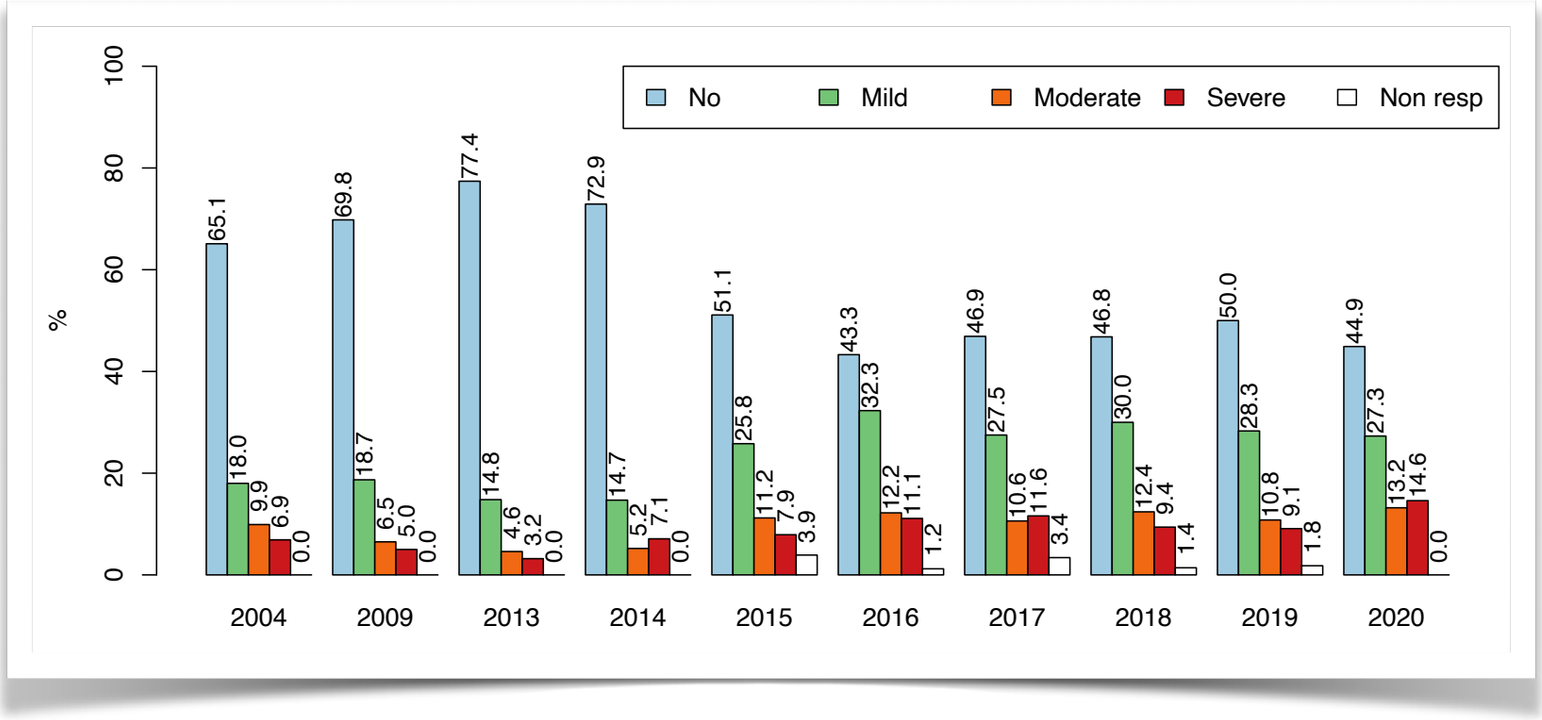
The dynamical model for this sub-network is

$$FoodSecurity_{1t} = \theta_{01,t} + \theta_{11,t} HIncome_t + \theta_{21,t} CFood_t + \epsilon_{1t}$$

$$FoodSecurity_{2t} = \theta_{01,t} + \theta_{12,t} HIncome_t + \theta_{22,t} CFood_t + \epsilon_{2t}$$

Food security in Brazil

- Food security measure (EBIA): it consists of 14 questions related to the direct experience of food insecurity. Score: 0 = food security; 1 to 3 = mild food insecurity; 4 to 5 = moderate food insecutiry; and 6 to 8 = severe food insecurity.
- Data: **Brazilian National Household Sample Survey (PNAD, 2004, 2009, 2013)**, **FAO (2014, 2015, 2016, 2017, 2018, 2019)**, **POF (2018)**, **Food for Justice (2020), Vigisan/Penssan (2020)**.



PNAD: Pesquisa Nacional por Amostra de Domicílios

POF: Pesquisa de Orçamentos Familiares

FIES: Food Insecurity Experienced Scale

FAO: United Nations' Food and Agriculture Organization

Vigisan/Penssan: Rede Brasileira de Pesquisa em Soberania e Segurança Alimentar (Rede PENSSAN), como parte do projeto VigiSAN (Vigilância da Segurança Alimentar e Nutricional)

Joint work with Luiz Eduardo S. Gomes (PhD student, IM, UFRJ).

Ongoing work

- In the case of **discrete data**,

$$Y_{it} \mid Y_{\Pi_i} = j \sim Multinomial(M_{ij}, \theta_{ijt})$$

- We extend the static model (Heckerman, 1995) to account for estimation of probabilities evolving smoothly over time (Fonseca and Ferreira, 2017), allowing for detection of **change of regimes**, **sustainability of policies** etc.

$$\mathbf{Y}_{it} \perp \mathbf{Y}^{t}_{Q_i} \mid \mathbf{Y}^{t}_{\Pi_i}, \mathbf{Y}^{t-1}_{i},\ i = 1, \ldots, p,$$

$$\boldsymbol{\theta}_{ijt} = \frac{1}{S_{ijt}} \boldsymbol{\psi}_{ijt} \odot \boldsymbol{\theta}_{ij,t-1},\ \psi_{ijk,t} \sim Beta(\delta_{ij} a^{(t-1)}_{ijk}, (1-\delta_{ij}) a^{(t-1)}_{ijk})$$

- Data: **Brazilian National Household Sample Survey (PNAD)**, consisting of 14 questions related to the direct experience of food insecurity. Score: 0 = food security; 1 to 3 = mild food insecurity; 4 to 5 = moderate food insecutiry; and 6 to 8 = severe food insecurity.

More about Dirichlet evolution: *T.C.O. Fonseca and M.A.R. Ferreira (**2017**) Dynamic Multiscale Spatiotemporal Models for Poisson Data (2017), Journal of the American Statistical Association 112:517, 215-234*

Graphical decision models via quantile regression

Joint work with: Kelly C Gonçalves (DME-UFRJ), Guilherme Oliveira (CEFET-MG), Victoria Silveira (undergraduate student IM-UFRJ) and Family Frias (undergraduate student IM-UFRJ)

- Main goal: extend the Bayesian Network methodology to account for ***robust solutions***.

- In particular, our project focus on three main areas: Food security, Gender inequality, and Birth outcomes in the Brazilian health system.

- This proposal's essential characteristic is collaborating closely with other researchers from different backgrounds, such as epidemiologists, economists, nurses, social scientists, etc.

- We give particular attention to the final usability of our modeling framework. Thus, we are developing a shiny app to allow practitioners to compute risks and compare decisions.

References

A. Hanea, M. McBride, M. Burgman, B. Wintle, F. Fidler, L. Flander, C. Twardy, B. Manning and S. Mascaro, (2017). Investigate discuss estimate aggregate for structured expert judgement. Int. J. Forecast. 33 (1), 267–279.

D. Heckerman et al. (1995), Learning Bayesian Networks: The Combination of Knowledge and Statistical Data, Machine learning 20, 197-243.

U.B. Kjærulff and A.L. Madsen (2013), Bayesian Networks and Influence Diagrams: A Guide to Construction and Analysis, Springer.

A. O'Hagan, C. E. Buck, A. Daneshkhah, J. R. Eiser, P. H. Garthwaite, D. J. Jenkinson, J. E. Oakley and T. Rakow (2006), Uncertain Judgements: Eliciting Experts' Probabilities, John Wiley & Sons, Ltd.

J. Oakley, A. O'Hagan (2016). SHELF: Tools to support the sheffield elicitation framework. http://tonyohagan.co.uk/shelf.

C. M. Queen and J. Q. Smith (1993) Multiregression dynamic models.Journal of the Royal Statistical Society. Series B (Methodological),55, 849–870.

Acknowledgments

- CNPq grant 04/2021
- CNPq grant CNPq/MCTI/FNDCT 18/2021
- Faperj grant E_27/2021 - APQ1
- EPSRC-funded GCRF Accelerator Account Fund 2019-2020
- Posdoctorade fellowship (AS&RU, Warwick University) 2019-20120
- National Lottery Heritage Fund (EPSRC grant EP/R511808/1)