17th World Conference on Earthquake Engineering, 17WCEE
Sendai, Japan - September 13th to 18th 2020

BAYESIAN UPDATING OF EXPOSURE MODEL INFERENCES USING INSPECTION DATA

M. Palmieri[1], A. Christodoulou[2], D.N. Grant[3]

[1] Senior Engineer, Arup (Netherlands) Michele.Palmieri@Arup.com
[2] Senior Engineer, Arup (Netherlands) Alex.Christodoulou@Arup.com
[3] Associate Director, Arup (ADE,UK) Damian.Grant@Arup.com

Abstract

This paper will present a methodology developed for the Bayesian updating of an inventory exposure model using data collected from inspection of a small subset of the inventory. The methodology was developed in the context of a regional exposure model used for characterising the seismic risk caused by induced earthquakes in the Groningen region. The initial (prior) model for the assignment of buildings in the inventory into building typologies was based on expert-judgement-based inference rules that assigned a probability to each of a range of different building typologies based on construction period and other relevant features of the dataset. As the study developed, actual building information from on-site inspections became available, which presented a potential for an improvement of the initial inferences. Given that inspection data rarely covered the full range of construction periods and typologies, there was a need for the development of a methodology to perform updates to the inferences and to the model without large building-by-building variation across adjacent years. The developed methodology presented a rational and consistent probabilistic basis for incorporating the inspection data while retaining some of the inferential value from the expert-judgement based assignment.

Keywords: Seismic Risk, Exposure Model, Bayesian inferences, Data-based inferences

1. Introduction

An accurate exposure model is a key component of seismic risk assessment and is combined with hazard and vulnerability models to estimate seismic losses over a region or portfolio. An exposure model should identify the location of buildings (and possibly other structures), assign building typologies that are linked to fragility and vulnerability models, and quantify the number of occupants expected at different times of the day. Depending on the end use of the risk assessment model and the data available, exposure can be aggregated at different geographical scales (often corresponding to civic administrative units or postal codes), or on a building-by-building basis.

Various data sources can be combined to develop an exposure model for a region, including municipal or national databases, census data, automated or manual inferences based on satellite imagery or street-level images (such as Google Streetview), or by carrying out individual building assessments (possibly aided by the availability of structural drawings). The latter is probably the most reliable (particularly if data is collected specifically according to the building taxonomy used in the risk assessment) but can also be time consuming over a large exposure model. In the absence of reliable data about building typologies, the last recourse is to expert judgement, based on any available information such as construction year, building size and usage. Depending on the homogeneity of buildings in the study area (and the quality of the experts!), this latter approach can be subject to significant uncertainty.

In this paper, some developments relating to the construction of a regional exposure model used for characterising seismic risk caused by induced earthquakes in the region around the Groningen gas field in the Netherlands are discussed. This paper describes the Bayesian approach that was used to consider newly collected data to update building typology assignments in the database. The next section gives an overview of the project, with a focus on the exposure model and the available data sources that were used in its development. Following sections describe the Bayesian inferences that were made to adjust the initial model based on new data.

2. Project Background

2.1 Induced seismicity risk model

A seismic risk assessment is being carried out for induced seismicity in the Groningen region in the north of the Netherlands, to investigate the "local personal risk" for occupied buildings. The regional seismic risk model comprises five main components [1]:

1. The seismicity model, which takes into account the expected number, magnitude, mechanism and location of earthquakes occurring in the area, as a function of the seismic moment induced by gas extraction in the field;
2. Ground motion prediction equations (GMPEs), which estimate the spectral accelerations caused by different earthquake scenarios at different locations in the area;
3. The exposure model, which provides the location and type of each of around 260,000 buildings in and around the gas field;
4. Fragility functions, which capture the probability and mode of building collapse given building type and the acceleration and duration of an earthquake ground motion [2] [3].
5. The consequence model, which calculates the probability of fatality given the mode of building collapse and collapse type [2].

2.2 Exposure model

Buildings in the regional exposure model are characterised by structural typologies consisting of nine building attributes: the main geometric class [4], the main structural material and lateral load resisting system in the primary and secondary direction of the building, the presence of external walls, the floor material, the number of floors and the irregularity class. The last eight parameters were defined according to the Building Taxonomy developed by the GEM foundation [5].

A number of public databases were available to collect information about Groningen building stock. Furthermore, other characteristics, such as the geometric layout class and number of storeys, could be inferred with high confidence based on geometric properties. The extraction of width, length and height data given the existing open height map and building footprint data available in the region is described in more detail on [7]. The following geometric layouts were defined: 1) S (shed) 2) U (unit) 3) B (block) 4) W (barn-warehouse) 5) T (tower). Presence of irregularities in plan or elevation was inferred by analysis of street imagery [6].

Unfortunately, the construction material and structural system were not available in any of these databases, and these are the primary data used to assign fragility functions. These attributes were grouped into the "structural system" class. Until building inspection is carried out for a specific building or other reliable data sources are obtained, the assignment of structural system is inherently uncertain; therefore, in the exposure model, each building is given a probability of being a particular structural system, based on the most likely system retrieved for inspected building within similar geometric layout. For example, a building may be assigned 60% probability of being an unreinforced masonry (URM) cavity wall building with timber floors, and 40% probability of being URM cavity walls with concrete floors. In the final risk calculation, the fragility of each of these possibilities is weighted to obtain an overall risk, including this uncertainty.

The classification of geometric layout, structural system, storey count and irregularity class are outlined in Fig. 1.

2.3 Expert-judgement based prior distribution

Initial assignments of structural systems and their probabilities for each building were made using expert judgement. For 260,000 buildings, assessing each building (even as a line item on a spreadsheet) on an individual basis would have been both prohibitively expensive and difficult to maintain consistency. Therefore, inference rules were defined based on construction year, geometric layout and structural systems (which were all either directly available from primary databases or were able to be inferred from primary data). The date ranges and probabilities considered in the inference rules were developed based on meetings and online surveys of engineers with experience in construction or building surveys in the Groningen area. These inference rules were then applied to the buildings to determine initial assignments – or, in the language of Bayesian inference, a prior distribution.

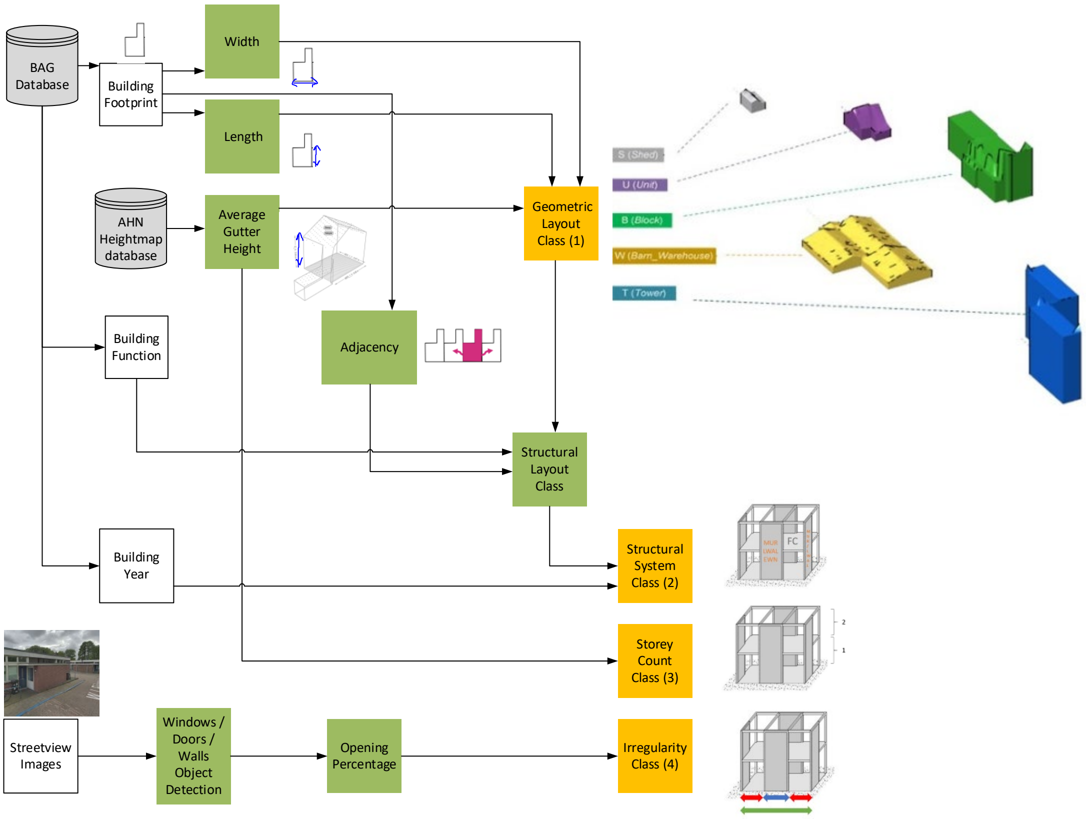

Fig. 1 – The inputs for the four independent classifications required for the case study project (in yellow). Calculated classification parameters are colour-coded in green, while externally acquired classification inputs are colour-coded in white.

2.4 Building inspection and Bayesian updating

As the study progressed, actual building information from on-site inspections became available, which presented a potential for an improvement of the initial inferences. Buildings that had been inspected were assigned to the relevant building typology with 100% confidence (and inference-based assignments were overwritten). The collected data also provided information about buildings that had not been inspected, since proportions of different typologies observed in the inspected buildings may be expected to be reflected in the non-inspected ones. The following sections describes the Bayesian approach that was used to update the initial exposure model to reflect this new data.

3. Bayesian Updating of Building Assignment Probabilities

3.1 Data extraction and formatting

For the inspected buildings, the building construction year and structural system assignment were aggregated. An example set of 47 inspected buildings is shown in Table 1. The taxonomy strings in the "Structural system" column are based on an extract of the GEM taxonomy that was used on this project (the specific meaning of each string is not important for the scope of this paper; hereafter, only the reference number will be referred to). The numbers in the other cells represent the number of buildings observed with that construction year and structural system. Based on this relatively small sample, we could infer that buildings with the structural system represented by the 2nd taxonomy string ("MUR/LWAL/MUR/LWAL/EW/FC") are the most common over this construction period.

Table 1. Example aggregation and reformatting of inspection data

Ref#	**Structural system**	**1977**	**1978**	**1979**	**1980**	**1981**	**1982**	**1983**
1	MUR/LWAL/MUR/LWAL/EWN/FW	0	0	0	1	0	0	0
2	MUR/LWAL/MUR/LWAL/EW/FC	4	16	2	7	2	1	3
3	W/LWAL/W/LWAL/EWN/FW	2	1	0	1	0	0	0
4	MUR/LWAL/MUR/LWAL/EW/FW	2	1	2	1	0	0	0
5	CR+CIP/LWAL/CR+CIP/LWAL/EWN/FC	1	0	0	0	0	0	0

3.2 Introduction of a smoothing operation

If a very large dataset of inspected buildings were available, it may be possible to normalise each column of Table 1, and use this data directly for subsequent inferences on un-inspected buildings. Table 2 shows the result of this operation. In this case, one would infer that a building constructed in 1978 is very likely (89%) to be Structural System 2, whereas if it was built in 1979, it is 50% likely to be system 2 and 50% likely to be system 4. Furthermore, in some years where even fewer buildings were observed (e.g. 1981, 1982, 1983), one would infer 100% probability of being assigned to a specific structural system. If each inspection was to influence only the inference for the exact building year of each inspected building, and not the inferences for the adjacent building years, then the inferences would be largely different for consecutive building years. This is unrealistic as there is likely to be a strong correlation between successive years in the types of structural system being building (with the exception of a few discrete step changes in construction practice brought about by disruptive events such as the introduction of new legislation or World War 2). For example, the inspection data shown in Table 2 would result in the probabilistic inferences for relative proportions shown in Fig. 2.

Table 2 Example resultant inferences if each inspection is considered only for 1 corresponding building year (& without any judgement-based prior inferences)

Structural System ref#	**1977**	**1978**	**1979**	**1980**	**1981**	**1982**	**1983**
1	0	0	0	0.1	0	0	0
2	0.44	0.89	0.5	0.7	1	1	1
3	0.22	0.06	0	0.1	0	0	0
4	0.22	0.06	0.5	0.1	0	0	0
5	0.12	0	0	0	0	0	0

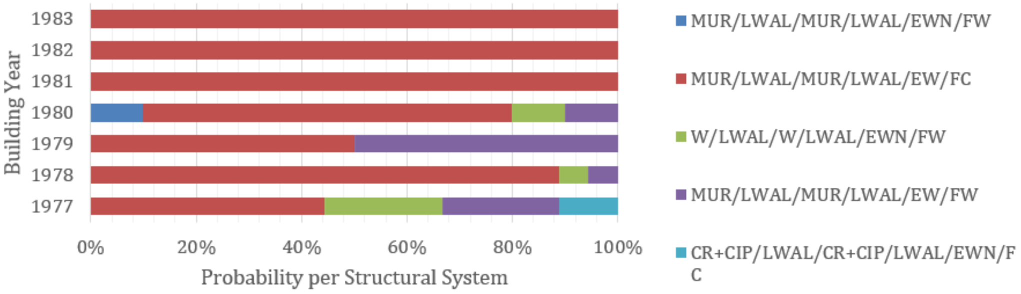

Fig. 2 – Visualisation of the direct data-driven inferences of Table 2

This can be avoided with the creation of synthetic data for years adjacent to the building year of the building inspected. This produces a similar outcome as the kernel density estimation methodology used in EDB V5 [4], where each inspection value was represented by a normal distribution kernel Eq.1.

$$(\Delta T) = \exp\left(-0.5\left(\frac{\Delta T}{T_0}\right)^2\right) \qquad (1)$$

The parameter T_0 can be interpreted as the bandwidth of the kernel. Fig. 3 shows an example with $T_0 = 10$ years.

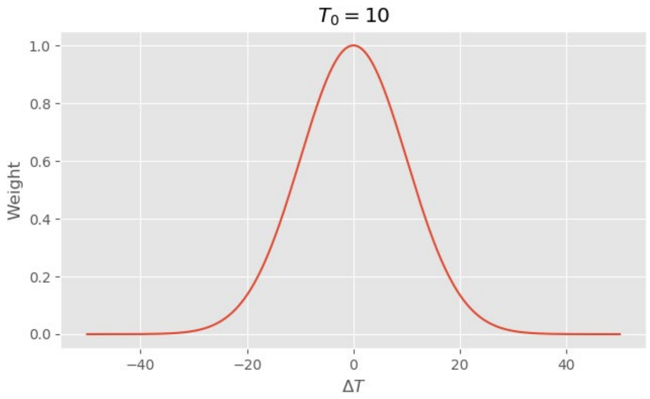

Fig. 3 – Gaussian correlation function when T0 = 10 years

Applying this correlation function to all the inspections of an inspection dataset (as for example the set shown in Table 2 results in a synthetic dataset, C, consisting of values, c_i, for each Structural System and building year. An example of the creation of such a synthetic dataset is shown in Table 3 given as input the inspection data of Table 2 and using a Gaussian correlation function with $T_0 = 1$ year.

Table 3 Example synthetic data created based on the inspection data of Table 2 with $T_0 = 1$ year.

Structural system	**1977**	**1978**	**1979**	**1980**	**1981**	**1982**	**1983**
MUR/LWAL/MUR/LWAL/EWN/FW	0.01	0.14	0.61	1.00	0.61	0.14	0.01
MUR/LWAL/MUR/LWAL/EW/FC	13.05	20.00	16.64	11.79	7.71	5.01	3.96
W/LWAL/W/LWAL/EWN/FW	2.62	2.35	1.48	1.16	0.62	0.14	0.01
MUR/LWAL/MUR/LWAL/EW/FW	2.89	3.56	3.48	2.37	0.89	0.16	0.01
CR+CIP/LWAL/CR+CIP/LWAL/EWN/FC	1.00	0.61	0.14	0.01	0.00	0.00	0.00

3.3 Bayesian inference update

In the final step, we update the prior Dirichlet probability distribution given the synthetic data created in the previous step.

Step 1 – Prior inference collection

The number of typologies present in the EDB V5 inferences of a given Structural Layout defines the number, r, of probabilities assigned to each Structural System for a given building year Eq. (2)

$$(p_1, p_2, \dots, p_r) \tag{2}$$

An example of such a probability vector can be seen in the following Table 4:

Table 4 Example vector of a probabilistic inference for the year 1980

Structural system	**1980**
MUR/LWAL/MUR/LWAL/EWN/FW	0.002
S/LFBR/S/LFBR/EWN/FC	0.011
S/LFM/S/LFM/EWN/FC	0.010
MUR/LWAL/MUR/LWAL/EW/FW	0.226
CR+CIP/LFM/CR+CIP/LFM/EWN/FC	0.006
CR+CIP/LWAL/CR+CIP/LWAL/EWN/FC	0.011
MUR/LWAL/MUR/LWAL/EW/FC	0.691
W/LWAL/W/LWAL/EW/FW	0.042

Step 2 – Identification of inspected Structural Systems not present in the k inferred systems

Structural Systems present in inspection data but not previously present in the inferences of EDB V5 are assigned $p_i = 0$, Eq. (3).

$$(p_1, p_2, \dots, p_r, p_{r+1}, \dots, p_n) \tag{3}$$

With $p_{r+1}, \dots, p_n = 0$.

If we import the inspection data from Table 2 then Table 4 would have to be modified as follows in Table 5:

Table 5 Modification of Table 4 to include inspected Structural Systems with 0 probability

Structural system	**1980 – prior inference**
MUR/LWAL/MUR/LWAL/EWN/FW	0.002
S/LFBR/S/LFBR/EWN/FC	0.011
S/LFM/S/LFM/EWN/FC	0.010
MUR/LWAL/MUR/LWAL/EW/FW	0.226
CR+CIP/LFM/CR+CIP/LFM/EWN/FC	0.006
CR+CIP/LWAL/CR+CIP/LWAL/EWN/FC	0.011
MUR/LWAL/MUR/LWAL/EW/FC	0.691
W/LWAL/W/LWAL/EW/FW	0.042
W/LWAL/W/LWAL/EWN/FW	0.000

Step 3 – Transformation of the prior inference to a Dirichlet probability distribution.

The *k* probabilities of Step 2 are now written as a Dirichlet distribution function in Eq. (4):

$$\mathrm{Dir}(\alpha_1, \alpha_2, \dots, \alpha_n) \tag{4}$$

With $\alpha_i = p_i\,\alpha_0$, where α_0 is the factor controlling the weighting given to the prior (EDB V5) inferences. Given that there is judgement involved in selecting an appropriate α_0 value, a sensitivity study will be presented in the following section to demonstrate the impact. (The Dirichlet distribution is an extension of the Binomial distribution to more than two possible outcomes.)

Step 4 – Calculation of the posterior probabilities

Subsequently, the α_i values (the weighting factor of each Structural System *i* for a given building year and Structural Layout) are added to the corresponding c_i value from the inspection-based synthetic dataset described in Section 3.2. The Dirichlet function giving the posterior probabilities for each Structural System in a given building year results in Eq. (5):

$$\mathrm{Dir}(\alpha_1 + c_1, \alpha_2 + c_2, \dots, \alpha_n + c_n) \tag{5}$$

Which results in the posterior probabilities of Eq. (6)

$$p|C = (p_{b1}, p_{b2}, \dots, p_{bn}) \tag{6}$$

The posterior probability p_{bi} for each Structural System *i* can be interpreted as a weighted sum of the prior probability p_{ai} weighted by α_i and the probabilities p_{ci} weighted by the synthetic count for each Structural System in each building year (Eq. (7)); i.e.:

$$p_{bi} = \frac{p_{ai}\,a_i + p_{ci}\,c_i}{a_i + c_i} \tag{7}$$

Using the data from Table 4 and Table 5 for this process would result in the second column for the posterior inference in

Table 6:

Table 6 Posterior inference using the example data from Tables 1 to 5.

Structural system	**1980 – prior inference**	**1980 – Posterior inference with α0 = 10, T0 = 1 year.**
MUR/LWAL/MUR/LWAL/EWN/FW	0.002	0.039
S/LFBR/S/LFBR/EWN/FC	0.011	0.004
S/LFM/S/LFM/EWN/FC	0.010	0.004
MUR/LWAL/MUR/LWAL/EW/FW	0.226	0.176
CR+CIP/LFM/CR+CIP/LFM/EWN/FC	0.006	0.002
CR+CIP/LWAL/CR+CIP/LWAL/EWN/FC	0.011	0.005
MUR/LWAL/MUR/LWAL/EW/FC	0.691	0.710
W/LWAL/W/LWAL/EW/FW	0.042	0.016
W/LWAL/W/LWAL/EWN/FW	0.000	0.044

Repeat steps 3.3.1 to 3.3.4 for all building years

The process described has to be repeated separately for every building year for which an inference needs to be developed. For example Fig. 4 illustrates how the inference shown in **Error! Reference source not found.** is updated if the inspections shown in Table 6 are inserted with $T_0 = 1$ year. As seen in this image, with a low bandwidth for T_0 the trends are still not captured as there is still a year-by-year inference variability that is contrary to the engineering-judgement based expectation. A higher T_0 is used in **Error! Reference source not found.**, where the engineering judgement is adjusted in a much more gradual manner.

One additional benefit of using the methodology is also apparent in Fig. 4: if inspections are only present on a specific period (as in this example where all imported inspection data was having a building year between 1977 and 1983), the inferences of years further away from the inspected buildings' years are not affected. Additionally, the transition between the years with updated inference and the years where the inference remains practically the same is gradual, as a result of the Gaussian kernel used whenever an inspection is imported.

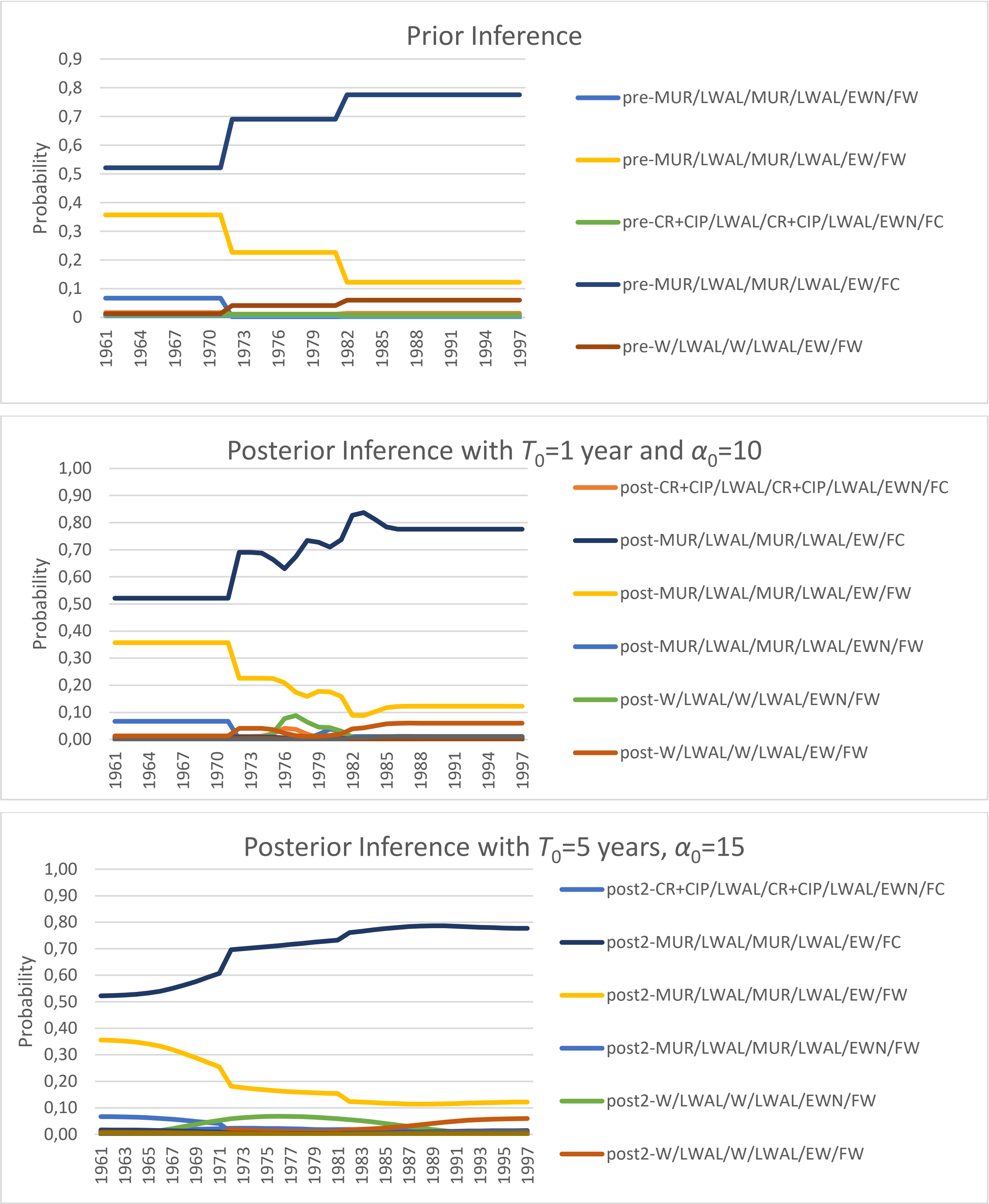

Fig. 4 – Prior inference, (Middle) Posterior inference for kernel width and weighting parameters T_0 = 1 years and α_0 = 10 (Down) Posterior inference for kernel width and weighting parameters T_0 = 5 years and α_0 = 15 . Using a wider gaussian kernel produces less year-by-year variability and reduces overfitting effects to the inspected data.

4. Methodology Calibration & Validation

The main scope developed for this assignment was to develop a method to evaluate the relative performance of the posterior inference in comparison to the prior inference.

To do this it is important to split the available inspected data into a training and a testing set to avoid recording increased performance only due to overfitting.

The optimisation objective developed for this study included testing the match between the count per class in the sample based on inspections and based on inferences (column *Class_Distance* in **Error! Reference source not found.**), as well as the count of buildings where the inspected class was also the most likely class (column *Ssy_1_matches* **Error! Reference source not found.**). In order to take both objectives into account at the same time, the value *Score_to_Maximize=Ssy_1_matches/class_distance* would have to be maximized. A few iterations of the optimisation algorithm are shown below in **Error! Reference source not found.** and Fig. 5:

Table 7 Data showing performance indicators (first 3 columns), given the calibration factors (last 2 columns)

Overall Score (to be maximized)	**SSY_1_Matches (to be maximized)**	**Class_Distance (to be minimized)**	**K (relative weighting to data over prior inferences)**	**T0 (Gaussian Kernel standard deviation)**
1.38	169	122.1	250000	10
1.40	169	120.8	250000	20
1.41	169	119.6	250000	40
1.19	143	119.7	250000	50
0.99	119	120.2	250000	70
0.99	119	120.1	250000	80
0.99	119	120.1	100000	80
0.99	136	136.8	100	80
0.98	136	138.9	100	40
1.39	169	121.9	1000	40
1.41	169	119.9	10000	40
1.41	169	119.5	100000	40

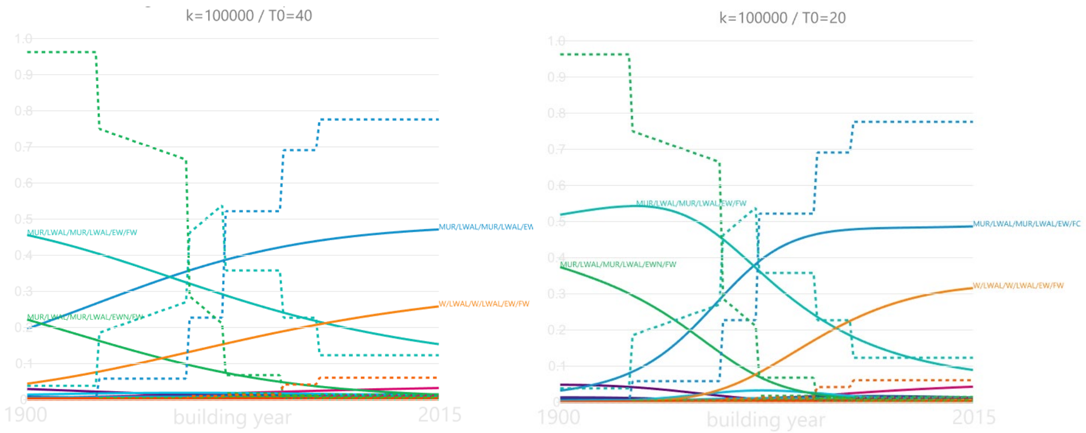

Fig. 5 – (Left) inference with wider Gaussian Kernel for the data-driven inference updates – performing better on the overall count match but worse for different year domains. (Right) Inference with thinner Gaussian Kernel for the data-driven inference updates – performing better when age brackets are tested separately. Dotted lines show the expert-based prior inferences.

5. Conclusions

An optimisation objective is hereby proposed which is a weighted combination between the overall match of the inferred class count and the number of buildings whose inspected structural system matches the most likely structural system. Further improvement of this methodology is the subdivision of the testing set in smaller building year domains to improve the performance of the inferences within age brackets.

If the main aim is to capture as well as possible the total counts of a sample, a wide Gaussian kernel seems to increase the matching rate (Class distance)

If a combined optimisation objective is chosen, considering also the number of times the most likely inferred class matches the inspected class, then an intermediate Gaussian kernel width value (T0=40) resulted in a better match.

Separating the testing dataset into buildings before 1960 and after 1960 further reduces the optimal T0 (Gaussian Kernel Width) from 40 to 20. This means that if the objective is to match the count as best as possible but with the probabilities coming as much as possible from buildings of similar building years then it is better to use a lower T0. A recommended study is to use even smaller building year domains to determine if the optimal is further reduced.

In all the variations ran, assigning negligible weight to the judgement-based inferences performed better. This should only be the case if we are certain that the learning and testing samples are unbiased – which is not the case in this study. Therefore, it is recommended that the study is repeated with a truly random sample within the class. If that is not performed, then the weighting to the prior inferences needs to be increased.

6. Acknowledgments

The authors gratefully acknowledge Nederlandse Aardolie Maatschappij BV for sponsoring the data collection, Helen Crowley for her precious feedback, and the and the many Arup colleagues who have contributed to the validation efforts and provided feedback for the implementation of updating inferences into the Exposure Model.

7. References

[1] Arup (2019): EDB V6 Rev A Cover Note. *Arup Report Number: 229746_031.0_NOT2055.*

[2] Crowley H. et al (2019): Report on the v6 Fragility and Consequence Models for the Groningen Field.

[3] Grant DN, Dennis J, Sturt R, McLennan D, Milan G, Huiu B, Palmieri M (2019), Explicit collapse prediction in the development of fragility functions for an unreinforced masonry building with non-linear finite element models, *SECED 2019 Conference*, London, UK.

[4] Arup (2017): Exposure Database V5 Tech Note. *Arup report number: 229746_052.0_AUX2008*

[5] Brzev, S., Scawthorn, C., Charleson, A. W., Allen, L., Greene, M., Jaiswal, K., & Silva, V. (2013): GEM Building Taxonomy (Version 2.0) (No. 2013-02). *GEM Foundation.*

[6] Ticinium Aerospace (2018): Automatic extraction, and measuring, of risk-related features in the Groningen (NL) area. *Arup report number: 229746_INC_031.0_AUX2110*

[7] Christodoulou et al. (2017): Automated building stock data mining and classification using open source data. *International Association for Shell and Spatial Structures symposium 2017*, Hamburg, Germany.