Centre for
Data Analytics

Insight

General Insurance Claims Modelling with Factor Collapsing and Bayesian Model Averaging

Sen HU, Dr Adrian O'Hagan, Prof Brendan Murphy

June 13, 2017

DCU NUI Galway OÉ Gaillimh UCC University College Cork, Ireland Coláiste na hOllscoile Corcaigh UCD DUBLIN SFI Science Foundation Ireland

Motivation:

- Model uncertainty with variable selection $\Rightarrow$ how confident we should be about the final model
- Existence of high multi-level factors - a factor having too many levels for a GLM structure $\Rightarrow$ model parsimony and interpretability issues
  - lack of sufficient number of observations
  - insignificant levels should be merged (too many parameters)

2 questions to answer:

- Which categorical predictors should be included in the model?
- Which categories within one categorical predictor should be distinguished?

Motivation

Factor collapsing (FC) assesses the optimal manner of categories: which differs from one another w.r.t dependent variable $\Rightarrow$ uncertainty about the optimal manner

Bayesian model averaging (BMA) takes such model uncertainty into consideration:

- variable selection uncertainty
- factor level selection uncertainty

Example: a question from "faraway" package [1]

Standard GLM output in R, for "Make" predictor in frequency model

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-1.812178	0.013758	-131.721	< 2e-16	***
...					
Make2	0.086384	0.021240	4.067	4.76e-05	***
Make3	-0.226013	0.025098	-9.005	< 2e-16	***
Make4	-0.640736	0.024196	-26.481	< 2e-16	***
Make5	0.161510	0.020235	7.982	1.44e-15	***
Make6	-0.331235	0.017375	-19.063	< 2e-16	***
Make7	-0.044705	0.023344	-1.915	0.0555	.
Make8	-0.008300	0.031606	-0.263	0.7929	
Make9	-0.069596	0.009956	-6.990	2.74e-12	***

Standard GLM output in R, for "Kilometres" predictor in severity model

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	8.39456	0.02335	359.474	< 2e-16	***
...					
Kilometres2	0.02455	0.01290	1.903	0.05718	.
Kilometres3	0.02124	0.01487	1.429	0.15327	
Kilometres4	0.04306	0.02073	2.077	0.03793	*
Kilometres5	0.03945	0.02205	1.789	0.07374	.

Factor collapsing

Set partition: grouping elements within a set into non-empty subsets, in such a way that every element is included in one and only one subsets. ("partitions" R package [2])

Partitioning 3-element set $\{1, 2, 3\}$:

$$\{\{1\}, \{2\}, \{3\}\}$$
$$\{\{1, 2\}, \{3\}\}$$
$$\{\{1, 3\}, \{2\}\}$$
$$\{\{1\}, \{2, 3\}\}$$
$$\{\{1, 2, 3\}\} \Rightarrow \text{variable removed}$$

Fit each (combination of) partition into a pre-specified model
Bell number increases nearly exponentially

BMA

Use BMA to average the best models (where possible)

$$Pr(\Delta|D) = \sum_{k=1}^{K} Pr(\Delta|M_k, D)Pr(M_k|D) \tag{1}$$

$$P(M_k|D) \approx \frac{\exp(-.5BIC_k)}{\sum_{r=0}^{K} \exp(-.5BIC_r)} \tag{2}$$

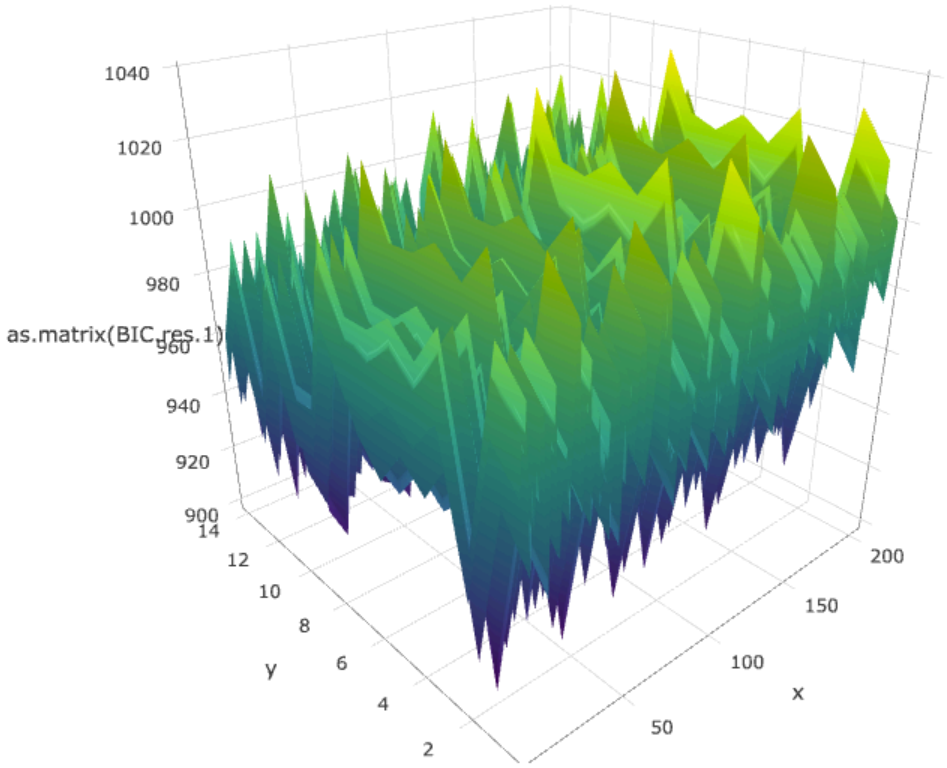
- Average over model prediction
- Average over model coefficients

Stochastic search

Number of set partitions increases nearly exponentially

$\Rightarrow$ computationally intensive

$\Rightarrow$ **it becomes an optimisation problem**



Simulated Annealing

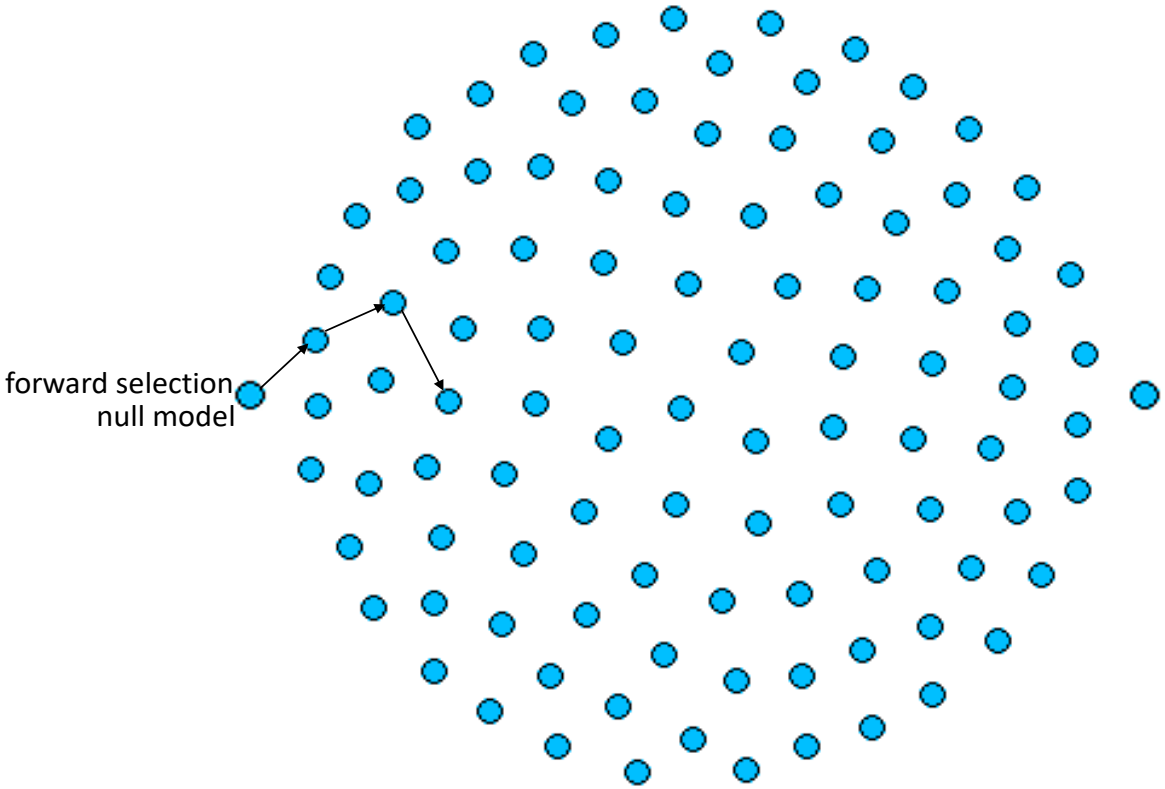
Global optimisation technique based on Monte Carlo method, similar to the MC^3 technique proposed in Hoeting et al. (1999) [3].

- Starting from a random state
- Make random state changes, accepting worse moves with probability determined by temperature
- Reduce temperature after reaching (close-to) equilibrium
- Stop once temperature gets very small

Other stochastic optimisation methods also work for this non-linear non-differentiable objective function, such as genetic algorithm etc.

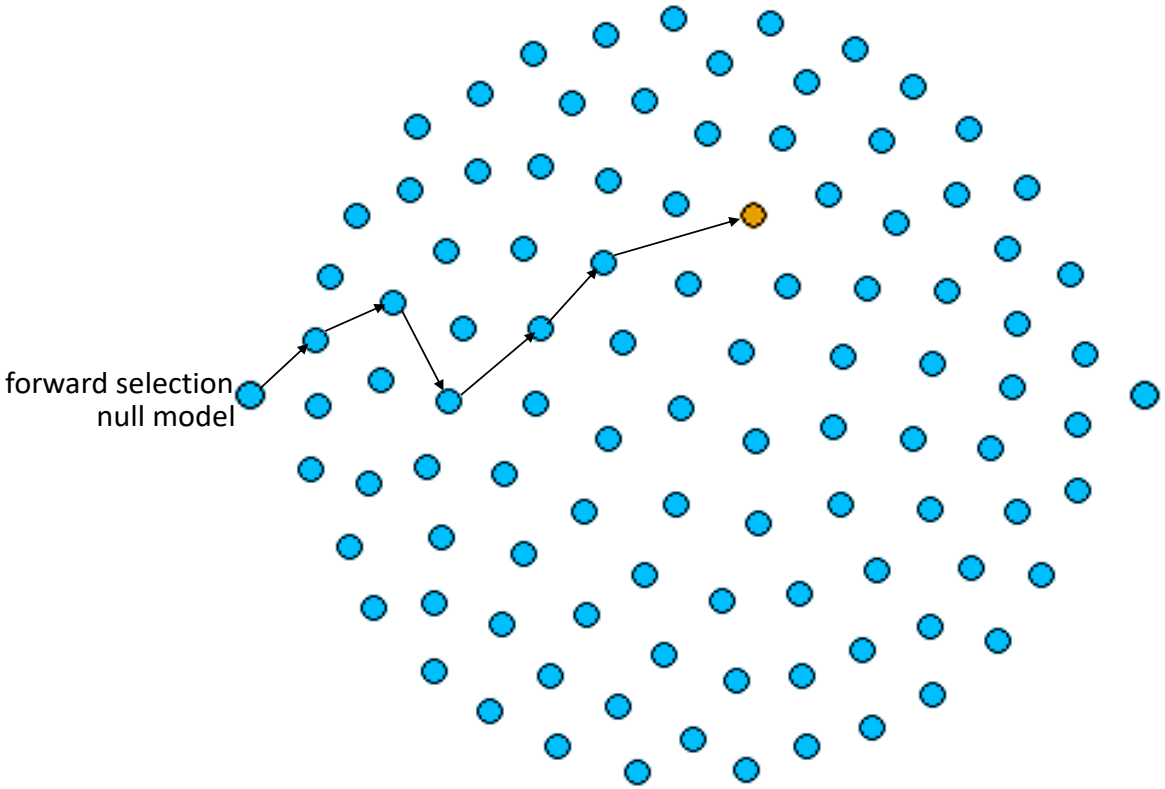
FC-BMA illustration

Comparing FC-BMA with stepwise selection using BIC/AIC:



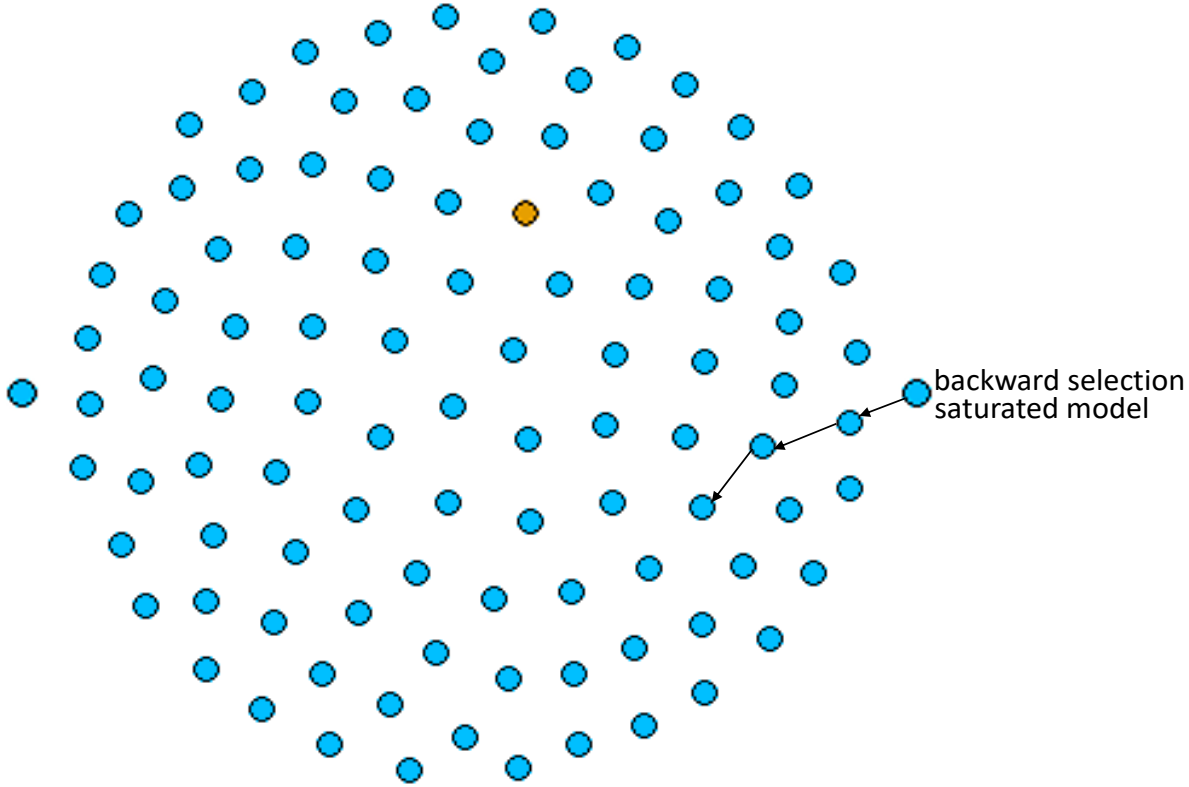
FC-BMA illustration

Comparing FC-BMA with stepwise selection using BIC/AIC:



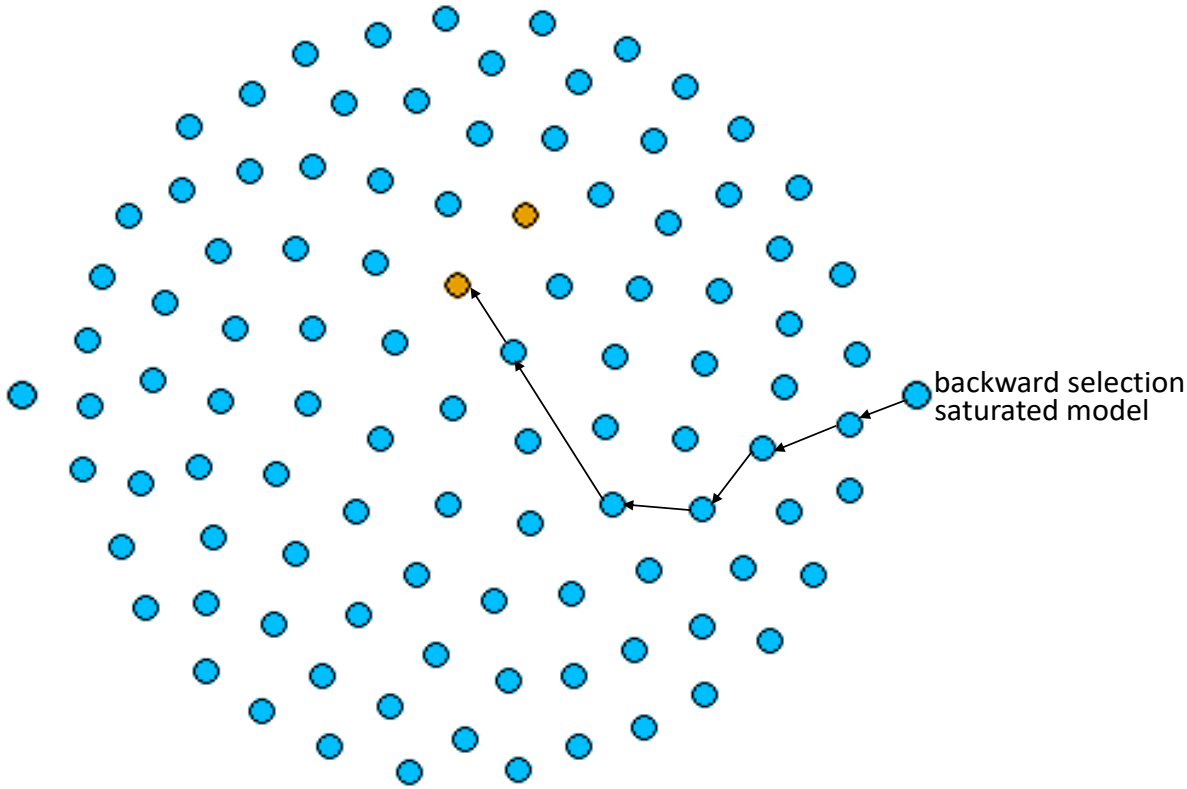
FC-BMA illustration

Comparing FC-BMA with stepwise selection using BIC/AIC:



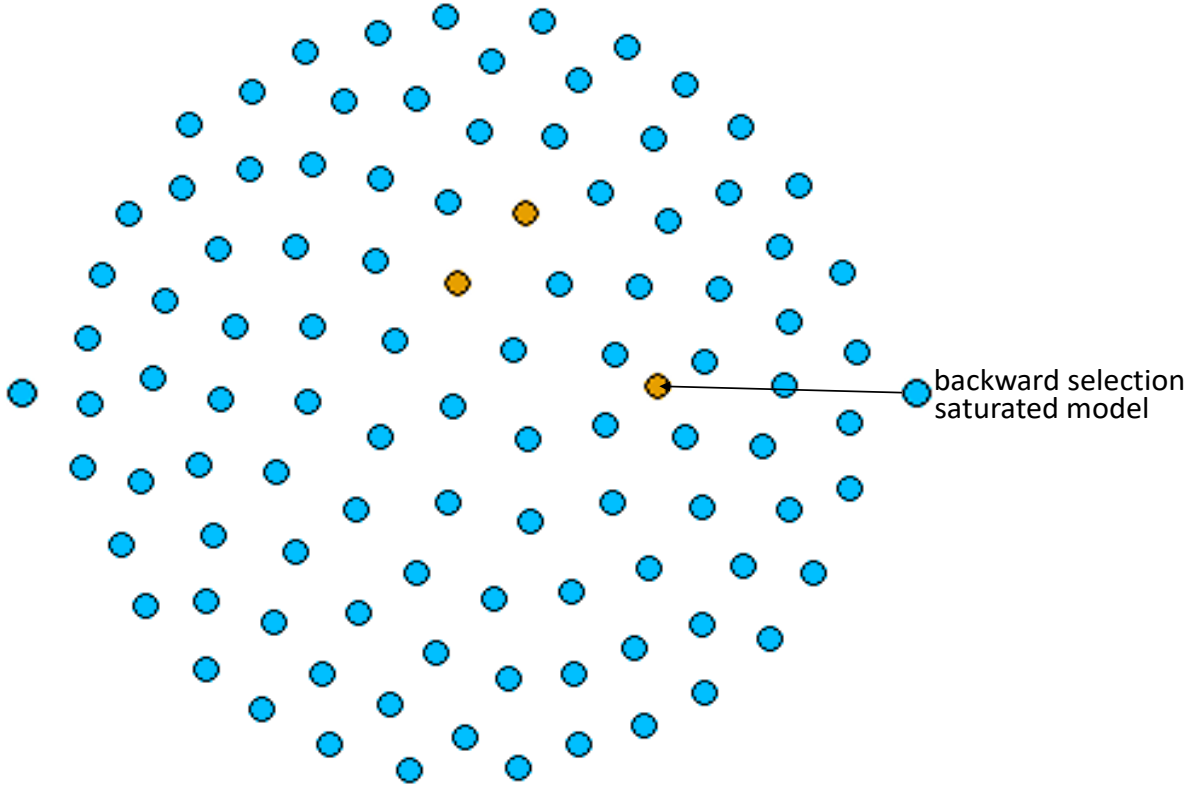
FC-BMA illustration

Comparing FC-BMA with stepwise selection using BIC/AIC:



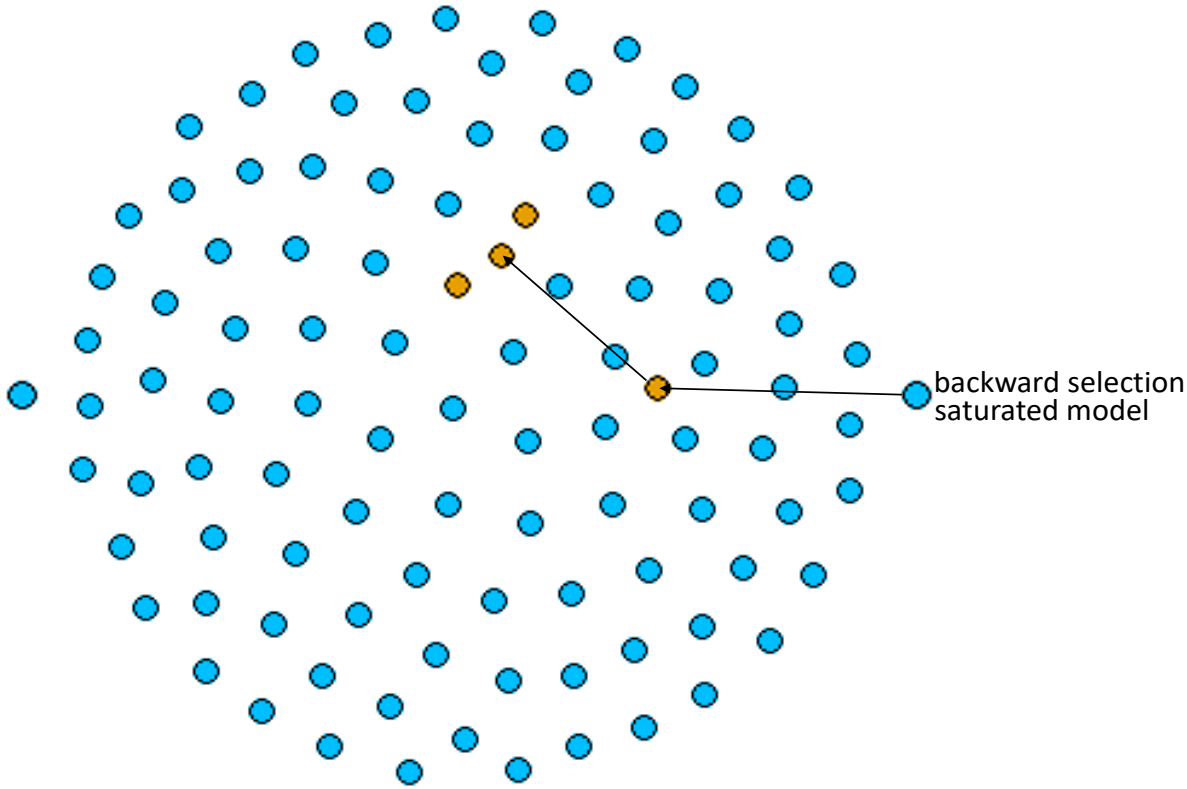
FC-BMA illustration

Comparing FC-BMA with stepwise selection using BIC/AIC:



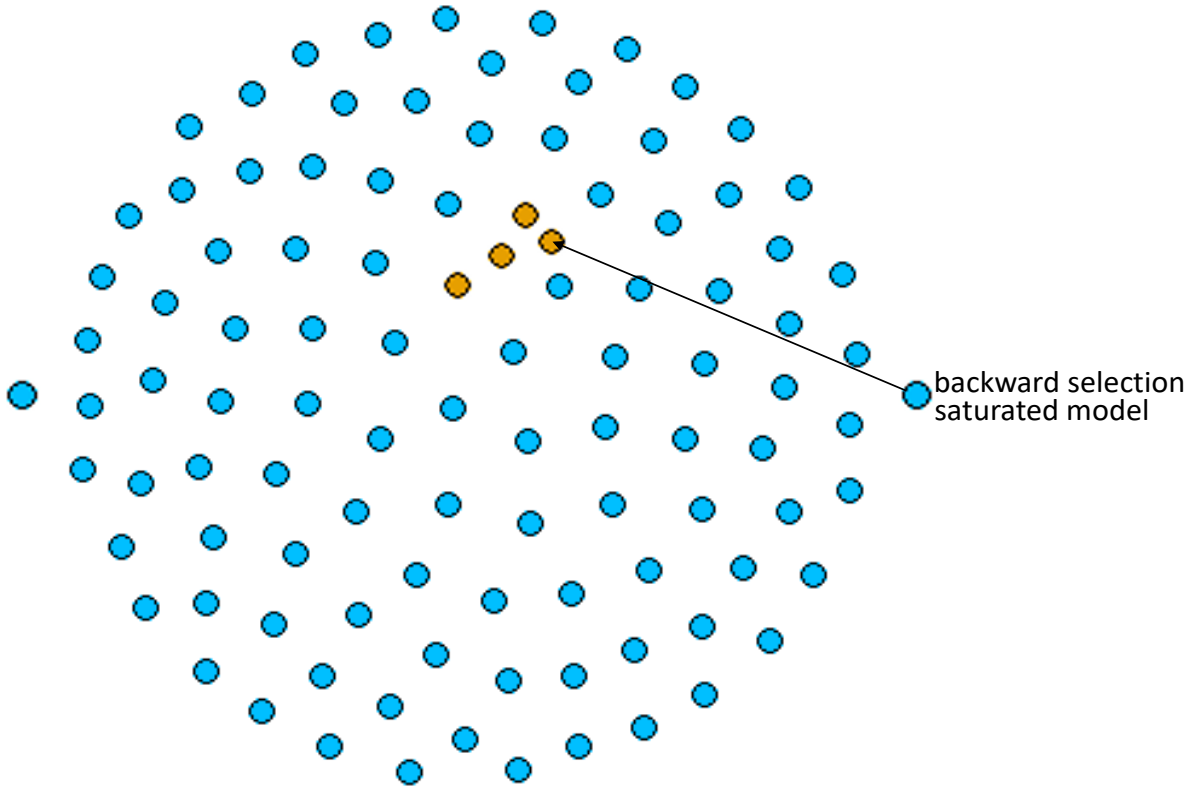
FC-BMA illustration

Comparing FC-BMA with stepwise selection using BIC/AIC:



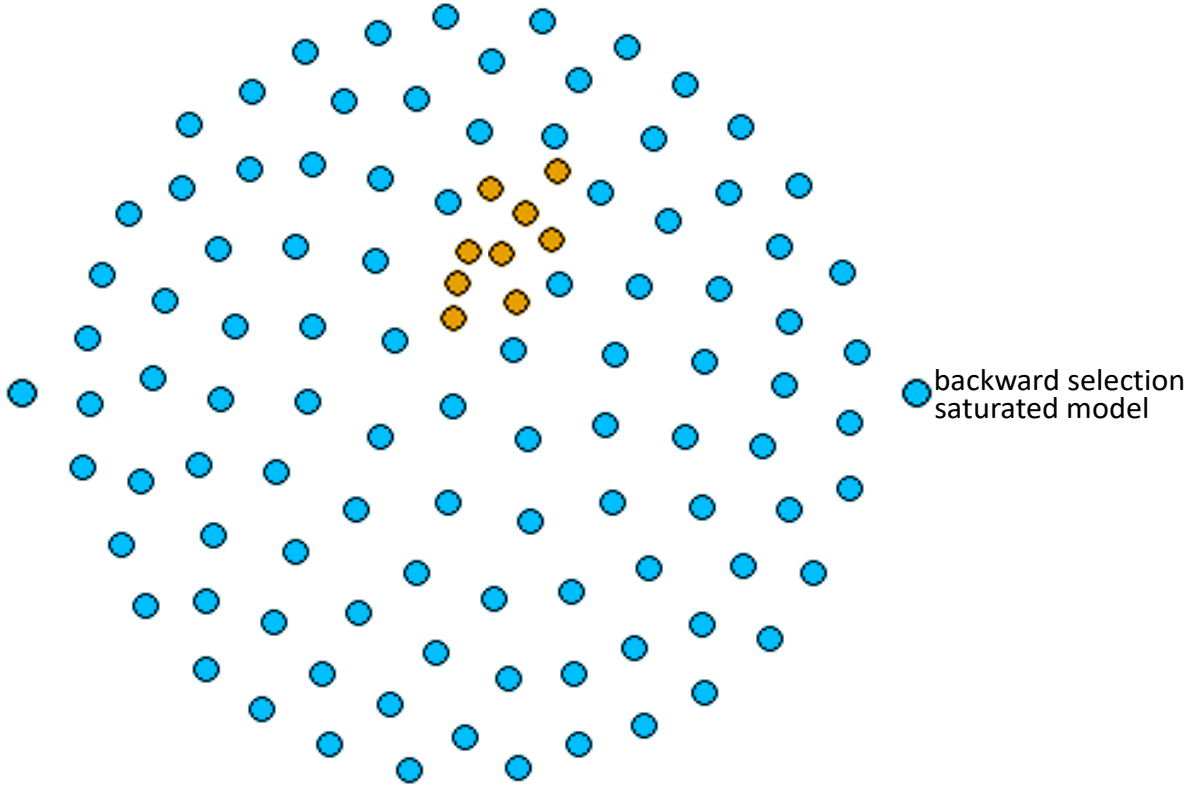
FC-BMA illustration

Comparing FC-BMA with stepwise selection using BIC/AIC:



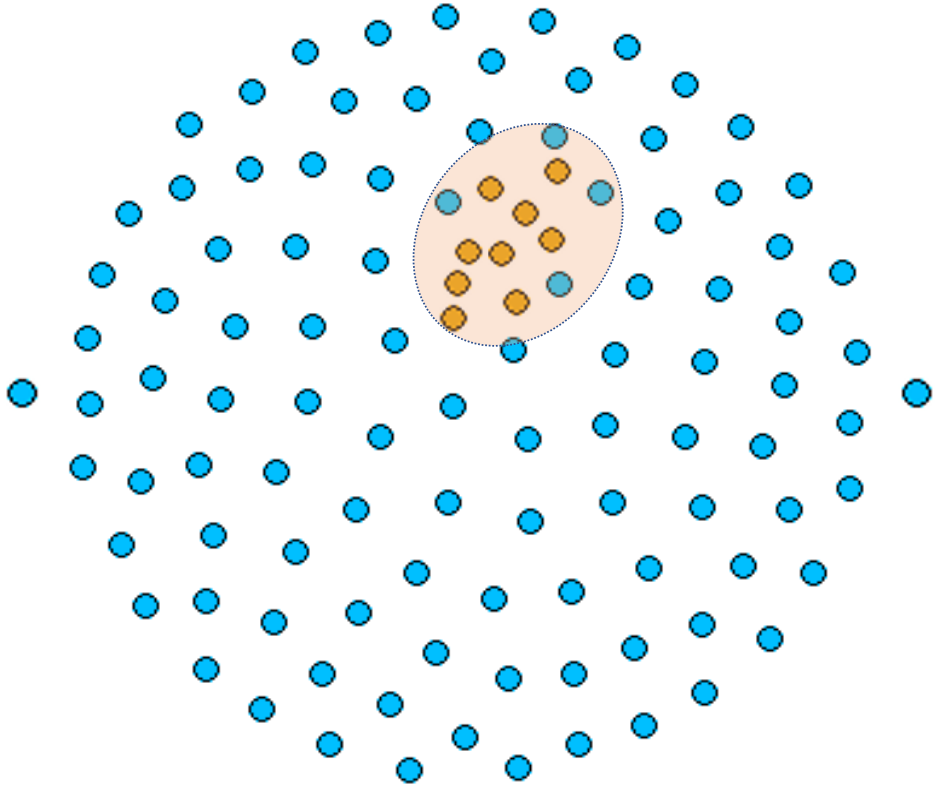
FC-BMA illustration

Comparing FC-BMA with stepwise selection using BIC/AIC:



FC-BMA illustration

Comparing FC-BMA with stepwise selection using BIC/AIC:



Following up the example...

Table: Results for collapsing "Make" factor only in frequency model. Here only the best 5 models (based on their BIC values) are shown.

Make: 1, 2, 3, 4, 5, 6, 7, 8, 9		
combination	BIC	BMA weight
(1,8)(2)(3)(4)(5)(6)(7,9)	10301.11	0.34579
(1,8)(2,5)(3)(4)(6)(7,9)	10301.81	0.24257
(1,7,8)(2)(3)(4)(5)(6)(9)	10303.44	0.10764
(1,7,8)(2,5)(3)(4)(6)(9)	10304.15	0.07541
(1)(2)(3)(4)(5)(6)(7,8,9)	10304.92	0.05136

Following up the example...

Table: Result for collapsing "Kilometres" factor only in severity model, only the best 5 models (based on BIC values) are shown.

Kilometres: 1, 2, 3, 4, 5		
combinations	BIC	BMA weight
(1)(23)(45)	1874293	0.90779
(1)(2)(3)(45)	1874299	0.05977
(1)(23)(4)(5)	1874300	0.03043
(1)(2)(3)(4)(5)	1874305	0.00200
(1)(25)(3)(4)	1874338	0.00000

Irish counties

Irish county level clustering with an Irish GI insurer:

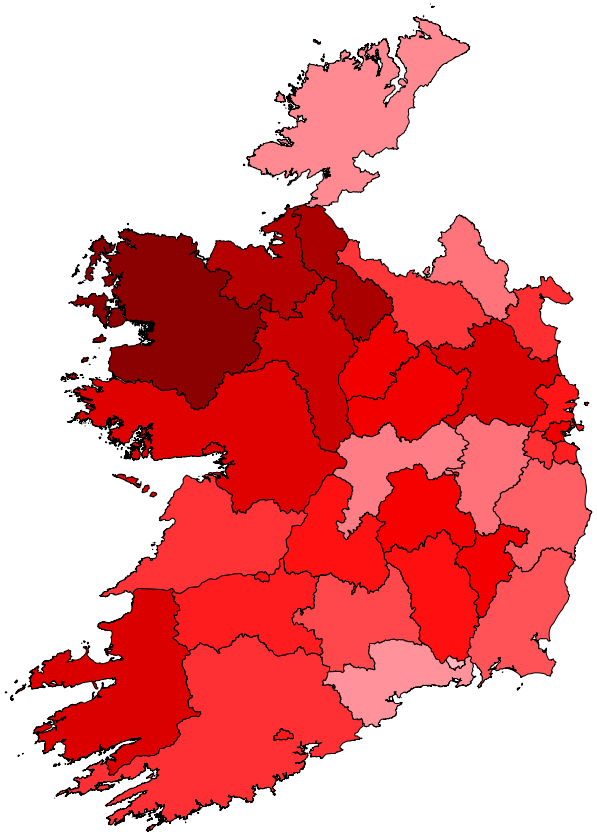


Figure: Frequency

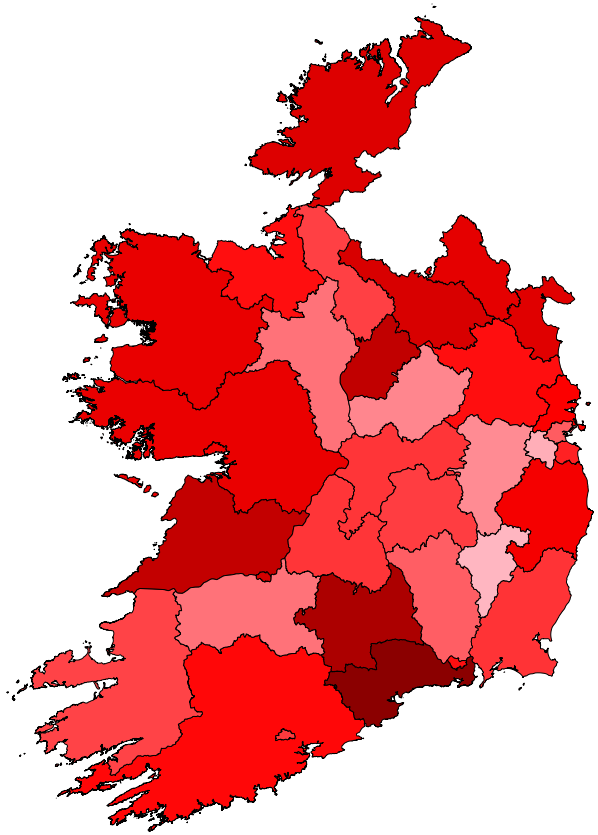


Figure: Severity

County	model coef.	new coef.
Waterford City	-6.6556	-6.6415
Unknown	-6.6130	-6.6415
Waterford County	-6.6073	-6.6415
Donegal County	-6.5959	-6.6415
Offaly County	-6.5787	-6.5733
Monaghan County	-6.5670	-6.5733
Kildare County	-6.5638	-6.5733
Wicklow County	-6.5397	-6.5733
Wexford County	-6.5217	-6.5733
South Tipperary	-6.5063	-6.5001
Cavan County	-6.4809	-6.5001
Clare County	-6.4764	-6.5001
Cork County	-6.4738	-6.5001
Louth County	-6.4720	-6.5001
South Dublin	-6.4708	-6.5001
Dun Laoghaire-Rathdown	-6.4489	-6.4648
Limerick County	-6.4473	-6.4648
Cork City	-6.4385	-6.4648
Fingal	-6.4379	-6.4648
North Tipperary	-6.4323	-6.4648
Limerick City	-6.4306	-6.4648
Kilkenny County	-6.4299	-6.4648
Laois County	-6.3923	-6.3766
Carlow County	-6.3865	-6.3766
Longford County	-6.3813	-6.3766
Westmeath County	-6.3808	-6.3766
Dublin City	-6.3694	-6.3766
Galway City	-6.3421	-6.3766
Galway County	-6.3415	-6.3766
Kerry County	-6.3323	-6.3766
Meath County	-6.3282	-6.3766
Roscommon County	-6.3031	-6.3766
Sligo County	-6.2503	-6.2106
Leitrim County	-6.2282	-6.2106
Mayo County	-6.1615	-6.2106

Irish counties

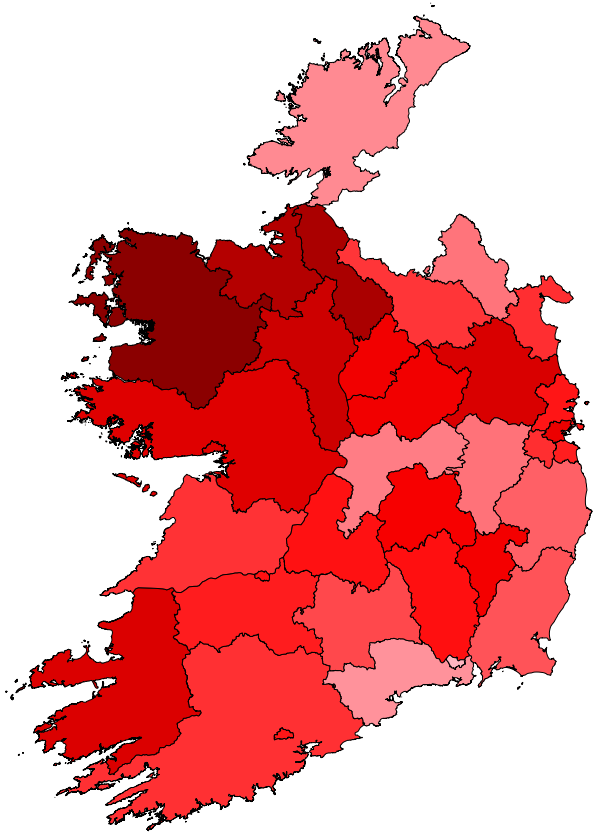


Figure: Frequency: before clustering

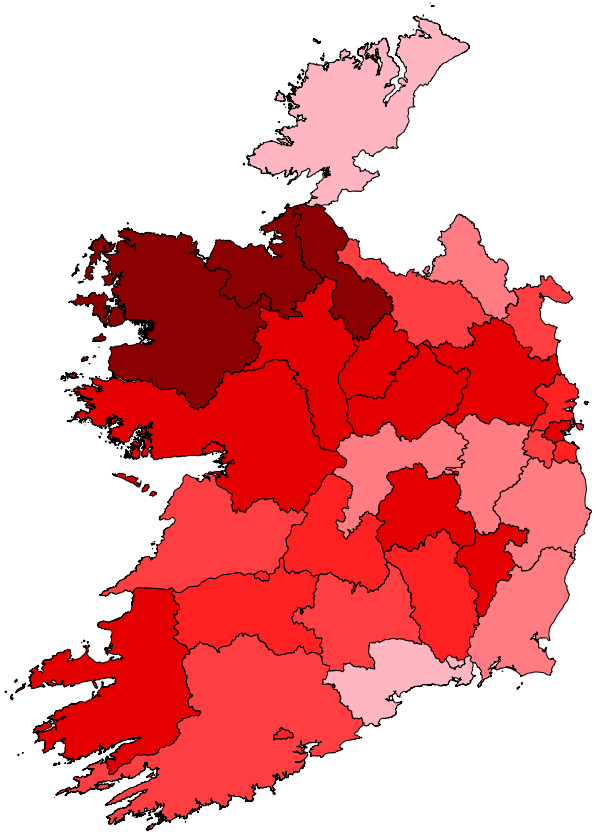


Figure: Frequency: after clustering

Table: (Subset of) Frequency model coefficients for the baseline standard GLM, and results of FC-BMA. Categorical levels are of increasing order based on the standard GLM. Only 5 are selected here for illustration.

	Std. GLM	BMA	Model 1	Model 2	Model 3	Model 4	Model 5
BIC			62807.2927	62807.3039	62807.3972	62807.4069	62807.4294
Model weights of all selected models			0.0233	0.0232	0.0221	0.0220	0.0218
Model weights of the 5 models			0.2074	0.2062	0.1968	0.1959	0.1937
Waterford City	-6.6556	-6.6359	-6.6414	-6.6399	-6.6326	-6.6341	-6.6311
Unknown	-6.6130	-6.6359	-6.6414	-6.6399	-6.6326	-6.6341	-6.6311
Waterford County	-6.6073	-6.6359	-6.6414	-6.6399	-6.6326	-6.6341	-6.6311
Donegal County	-6.5959	-6.6359	-6.6414	-6.6399	-6.6326	-6.6341	-6.6311
Offaly County	-6.5787	-6.6218	-6.5733	-6.6399	-6.6326	-6.6341	-6.6311
Monaghan County	-6.5670	-6.6080	-6.5733	-6.5732	-6.6326	-6.6341	-6.6311
Kildare County	-6.5638	-6.5695	-6.5733	-6.5732	-6.5689	-6.5674	-6.5645
Wicklow County	-6.5397	-6.5695	-6.5733	-6.5732	-6.5689	-6.5674	-6.5645
Wexford County	-6.5217	-6.5695	-6.5733	-6.5732	-6.5689	-6.5674	-6.5645
South Tipperary	-6.5062	-6.5263	-6.5000	-6.5023	-6.5006	-6.5674	-6.5645
Cavan County	-6.4809	-6.5004	-6.5000	-6.5023	-6.5006	-6.5011	-6.4980
Clare County	-6.4764	-6.5004	-6.5000	-6.5023	-6.5006	-6.5011	-6.4980
Cork County	-6.4738	-6.5004	-6.5000	-6.5023	-6.5006	-6.5011	-6.4980
Louth County	-6.4720	-6.5004	-6.5000	-6.5023	-6.5006	-6.5011	-6.4980
South Dublin	-6.4708	-6.5004	-6.5000	-6.5023	-6.5006	-6.5011	-6.4980

Table: Prediction comparison in Swedish TPML dataset, using MSE, Gini index, concordance correlation coefficient (CCC), Wasserstein distance, Kolmogorov-Smirnov test (KS-test), KL divergence respectively.

80% and 20% split		MSE	Gini	CCC	Wass.	KS-test	KL
Frequency	no FC-BMA	266.9408	0.8266	0.9968	3.0340	0.0736(0.3045)	0.0122
	FC-only	224.7803	0.8267	0.9943	2.9696	0.0788(0.2358)	0.0114
	FC-BMA(5)	456.3766	0.8267	0.9973	4.2012	0.0778(0.2535)	0.0113
Severity	no FC-BMA	14748455	0.0567	0.0409	1948.3340	0.4489(0)	0.2191
	FC-only	14664567	0.0576	0.0667	1825.0540	0.4067(0)	0.2178
	FC-BMA(5)	14666355	0.0576	0.0657	1822.9450	0.4033(0)	0.2178

Summary

- FC-BMA deals with model selection and uncertainty, categorical level selection simultaneously.
- It helps improve the model parsimony, interpretability, and prediction.
- Compared with other existing methods in literature, it does not require deciding extra parameters.
- It can be a challenge to obtain the optimum through stochastic optimisation, and may take a long time to reach the optimum.

References

- J. Faraway. "faraway: Function and datasets for books by Julian Faraway". In: *R package version 1.0.7* (2016).
- R. K. S. Hankin. "Additive integer partitions in R". In: *Journal of Statistical Software, Code Snippets* 16 (1 2006).
- Jennifer A Hoeting et al. "Bayesian Model Averaging: A Tutorial". In: *Statistical Science* 14.4 (1999), pp. 382–417. ISSN: 08834237.
- Torsten Hothorn, Frank Bretz, and Peter Westfall. "Simultaneous Inference in General Parametric Models". In: *Biometrical Journal* 50.3 (2008), pp. 346–363.

Q & A...