An Estimator of Winter Loss of Honey Bees in Sweden

Elenor Hagö

Bachelor's thesis
2018:K24

Lund University

Faculty of Science
Centre for Mathematical Sciences
Mathematical Statistics

An Estimator of Winter Loss of Honey Bees in Sweden

Bachelor Thesis in Mathematical Statistics
By Elenor Hagö
Supervised by Ullrika Sahlin, Centre of Environmental and Climate Research, Lund University

Spring 2018

Faculty of Science
Centre for Mathematics
Mathematical Statistics

LUNDS UNIVERSITET

Abstract

The aim of this thesis is to suggest a statistical model to estimate the loss rates of honey bees during winter in Sweden. The estimator is to be based on annual summary statistics, collected by the Swedish beekeeping organisation.

Regional specific estimates for winter loss rate is derived by a spatial and temporal hierarchical model with binomial response. The model is updated by Bayesian inference using Integrated Nested Laplace Approximations (INLA). Winter loss estimates were derived with probability intervals and presented on a map of Sweden.

The analysis shows that the average winter loss rates ranges between 10.2 and 19.7 % across the 21 regions, while the differences in average loss rates between years were increasing from ranges 9.3 to 18.3 % in 2015, 9.9% to 19.4 % in 2016 and 11.0 to 21.3 % in 2017. Regional differences were not linked to cultivation zones, when comparing different models with the information criterions WAIC and DIC.

The analysis included summary statistics from three years. It is possible to expand the model to include spatial and temporal interaction and trends over time by including summary statistics from more years. Estimates of winter loss rates based on data from several years have stronger properties compared to properties due to changes between regions, especially since the contributing beekeepers may vary a lot from year to year and in each region.

Acknowledgement

I would like to give a huge Thank you to Ullrika Sahlin for introducing me to this subject, her guidance and above all, her good wits.

Table of content

1. Introduction and aim.. 8
2. Method.. 8
   2.1 Data.. 8
      2.1.1 Counties of Sweden.. 9
      2.1.2 Cultivation zones.. 10
   2.2 Statistical model for colony loss rate.. 11
      2.2.1 Statistical models for spatial dependencies.. 12
      2.2.2 Statistical models... 13
      2.2.3 Model comparison by WAIC and DIC... 14
   2.3 Parameter estimation.. 14
      2.3.1 Bayesian inference using INLA... 15
   2.4 Implementation.. 17
3. Results….. 17
   3.1 Model 1... 17
   3.2 Model 2... 18
   3.3 Model 3... 19
   3.4 Model 4... 19
   3.5 Model comparison.. 20
      3.5.1 WAIC and DIC.. 20
      3.5.2 Regional loss rate.. 20
   3.6 Further results.. 21
      3.6.1 Regional loss rate.. 21
      3.6.2 Zone loss rate... 22
      3.6.3 Yearly loss rate.. 23
      3.6.4 Historical winter losses.. 24
      3.6.5 Goodness of fit... 25
4. Conclusions and discussion.. 26
5. References... 29

1. Introduction and aim

Winter loss in managed honey bees is the colonies that die during winter. Reliable estimates of the winter losses in managed honey bees can be used for monitoring and detecting trends in winter losses and how it is related to major drivers and stressors influencing honey bee health.

Every year, the Swedish beekeeping organisation collects data from their members on winter loss. Reporting is voluntary and about 50% of the members respond to the survey. The beekeeping organisation aggregates the number of colonies lost during winter in each region. The number of members reporting varies a lot between year and from different regions, which influences each aggregated data point.

I will derive probability intervals for loss rate with a bayesian model which has the property of giving the complete probability distribution of loss rate as a direct result. From that result a probability interval is summarised. This has otherwise been derived using confidence intervals of the log odds (Van der zee.R, et.al., 2012, p.32).

There is a possibility of spatial dependency between the different regions, meaning that colony loss in two or more regions is dependent on the same factors. Certain regions and cultivation areas are for example observed to have a shortage of pollen and nectar plants which has a negative effect on honey bees health (Jordbruksverket, 2009). Spatial variability will therefore be modeled as a fixed zone effect and a random spatial effect.

The aim of this thesis is to find a suitable estimator for loss rate during winter, based on this data. It will be achieved by the following objectives:

1) specify a statistical model for the data on loss rate during winter of honey bees in Sweden,
2) test if regional differences can be linked to cultivation zones
3) estimate historical winter losses in Sweden over the last three years.

2. Method

2.1 Data

The Swedish beekeeping organisation annually collects data and then presents it as summary statistics. The data has been aggregated on a regional level which is displayed as 25 regions and has been collected over the course of 3 years. The 25 regions were merged to 21, to represent the 21 counties of Sweden. There is a larger number of members then there is reporting members so the estimation of total loss, given all members, will be based on data from reporting members.

Computations will be performed in the software R using the R-INLA package (R Core Team, 2017).

2.1.1 Counties of Sweden

To give a sense of where Sweden's counties lie, a map of Sweden is presented below (figure 1). It is observed that a larger number of beekeepers report in the southern half of Sweden (figure 2). The total number of reporting members was 5551 in 2015, 6237 in 2016 and 6494 in 2017.

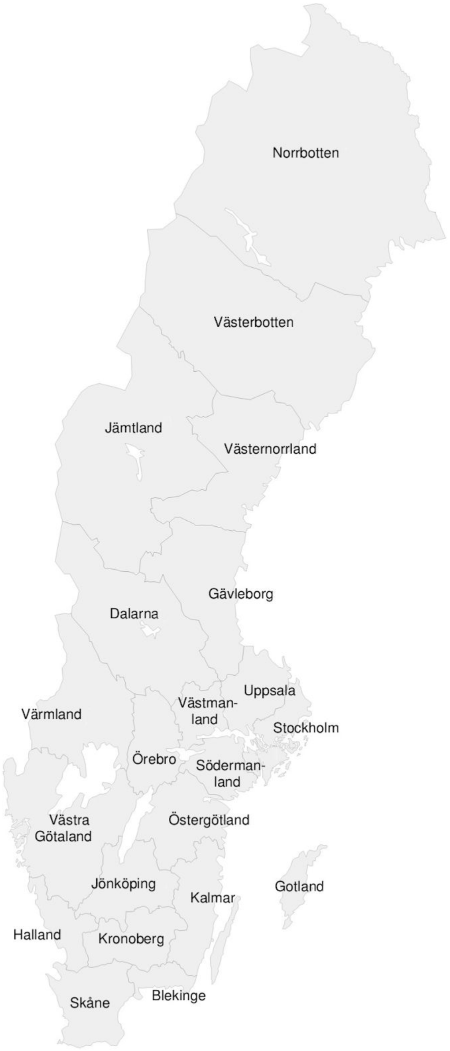


Figure 1. Counties of Sweden (Wikimedia Commons, 2018).

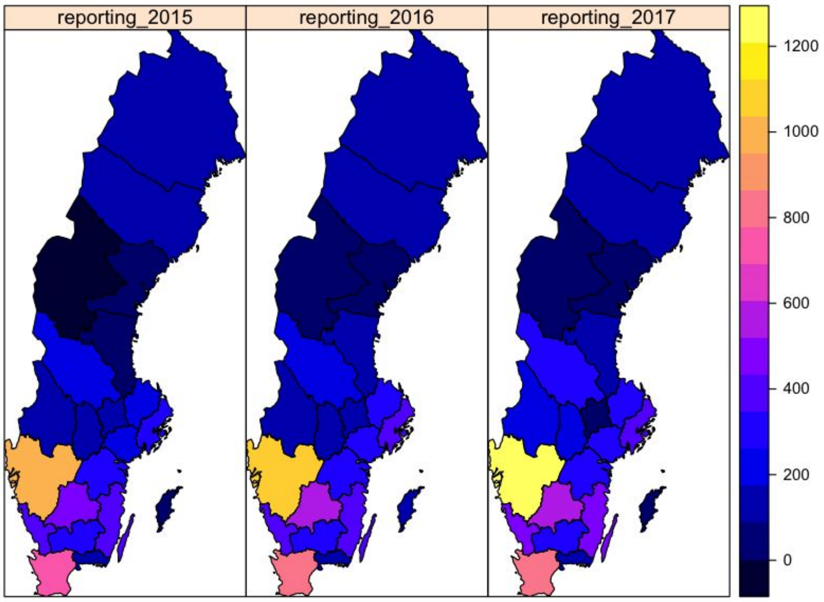


Figure 2. Number of reporting members in each region (2015-2017).

2.1.2 Cultivation zones

The counties of Sweden are divided into 5 zones according to a map of Sweden's cultivation zones (figure 3). Zone 1 is defined as cultivation zone 1, zone 2 as cultivation zones 2-3, zone 4 as cultivation zones 5-6 and zone 5 as cultivation zones 7-8. A list of which each county belongs to is also presented (table 1).

Table 1. Allocation of zone number to the swedish counties.

Region	Zone	Region	Zone	Region	Zone
Skåne	1	Västra Götaland	2	Värmland	4
Halland	1	Kalmar	3	Örebro	4
Blekinge	1	Jönköping	3	Dalarna	4
Gotland	1	Kronoberg	3	Västerbotten	5
Östergötland	2	Uppsala	4	Västernorrland	5
Södermanland	2	Gävleborg	4	Jämtland	5
Stockholm	2	Västmanland	4	Norrbotten	5

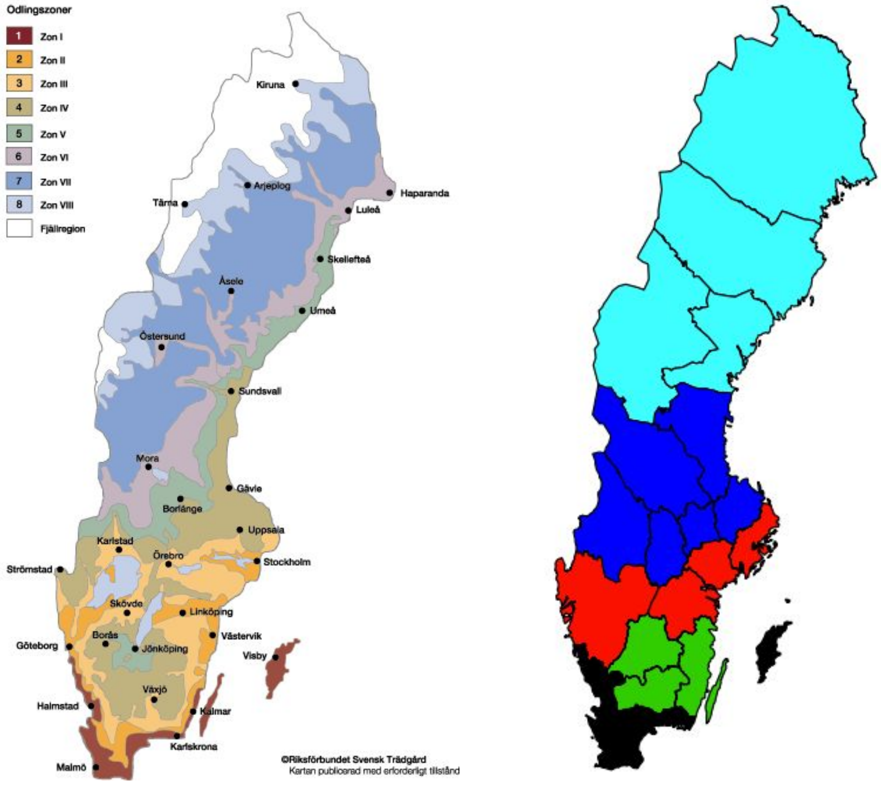


Figure 3. Left panel: The eight swedish cultivation zones (Riksförbundet Svensk trädgård, 2018). Right panel: Cultivation zones divided into 5 zones. Zone 1 is represented as black, zone 2 as red, zone 3 as green, zone 4 as dark blue and zone 5 as light blue.

2.2 Statistical model for colony loss rate

Colony loss rate is modelled by a generalized linear mixed model (GLMM), which consists of a family distribution for the response variable, a linear regression term with fixed and random effects and a link function. This type of model has previously been used for colony loss rate (Van der zee.R, et.al., 2012, p.25). The total colony loss for beekeeper *j* in region *i* year *t* is a binomial random variable under the assumption of the same loss rate for every colony. If Y_{jit} is the number of dead colonies after winter, n_{jit} is the number of colonies before winter and the parameter p_{it} is the loss rate of colonies, then total colony loss is defined as

$$(1) \qquad Y_{jit} \sim \text{Bin}(n_{jit}, p_{it})$$

Due to the fact that the data does not contain information about colony loss for all members but only for reporting members, certain (strong) assumptions will be made to be able to estimate the loss rate for the above model, regarding all members of the Swedish beekeeping organisation.

In this case, the data is aggregated across all beekeepers (who has reported) in a region. If assuming that the colony loss between beekeepers is independent and that all beekeepers in a region has the same probability for winter loss, then the following theorem can be applied.

Consider $X_1 \sim$ Bin(n, p) and $X_2 \sim$ Bin(m, p), independent binomial variables with the same probability p, then $X_1 + X_2$ is also a binomial variable with distribution Bin(n+m, p).

This results in the fact that the estimated loss rate with respect to reporting members would be the same as of that regarding all members.

(2) $$Y_{\cdot it} \sim \text{Bin}(n_{\cdot it}, p_{it})$$

To estimate p_{it}, a link function, η_{it}, in form of a logit transformation will be defined to remove any range restrictions, creating a probability between 0 and 1. (G. Rodriguez. 2007)

(3) $$\eta_{it} = logit\,(p_{it}) = log\,(\frac{p_{it}}{1 - p_{it}})$$

When estimated, the inverse logit retrieves the probability scale.

(4) $$p_{it} = logit^{-1}\,(\eta_{it}) = (\frac{e^{\eta_{it}}}{1 + e^{\eta_{it}}})$$

However, I want to estimate the loss rate with the assumption of both spatial and temporal differences and therefore will η_{it} be defined as a sum of regression terms. (Van der zee.R, et.al. 2012, p.25)

2.2.1 Statistical models for spatial dependencies

A bym model will be used to see if spatial dependencies can be observed. It is the union of an i.i.d model *v* and a besag model *u,* i.e. a random effect + spatial effect.

v_i is defined as the i.i.d model where the random variable $\mathbf{x} = (x_1, ..., x_n)$ is a vector of n counties that is independent and normally distributed with precision τ so that $x_i \mid \tau \sim N(0, \frac{1}{s_i \tau})$ where each x_i is conditionally independent, $s_i > 0$ is a fixed scale and i is the number of counties (R-INLA, 2018 a).

u_i is defined as the besag model for the random vector $\mathbf{x}$ so that $x_i \mid x_j,\ i \neq j, \tau \sim N(\frac{1}{n_i} \sum_{i \sim j} x_j, \frac{1}{n_i \tau})$ where n_i is the number of neighbours of county x_i and $i \sim j$

indicates that the two counties x_i and x_j are neighbours (R-INLA, 2018 b).

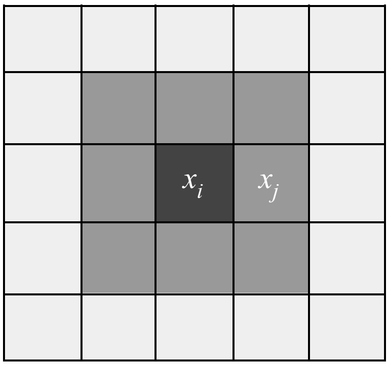


Figure 4. Example of neighbourhood structure. All medium gray squares are neighbours to the dark gray square, hence $n_i = 8$ in this illustration.

2.2.2 Statistical models

The following four models will be considered, which will include different combinations of terms. Model 1 is considered to be the null model.

Model 1 is defined as

(5) $\eta_{it} = b_0 + v_i + \delta_t$,

where b_0 is the intercept, v_i an unstructured spatial random effect and δ_t an unstructured temporal random effect.

Model 2 is defined as

(6) $\eta_{it} = z_i + v_i + \delta_t$,

where z_{it} is added to model 1 as a fixed effect for cultivation zones.

Model 3 is defined as

(7) $\eta_{it} = b_0 + v_i + u_i + \delta_t$,

where, compared to model 1, region is evaluated with an additional structured spatial random effect u_i.

Model 4 is defined as

(8) $\eta_{it} = z_i + v_i + u_i + \delta_t$,

where both, z_i from model 2 and u_i from model 3, are regarded as an addition to model 1.

The random effects v_i, u_i and δ_t are normally distributed with expectation zero and a variance that will be estimated and later presented as the precision $\tau_k = 1/\sigma_k^2$, k = v, u, δ .

2.2.3 Model comparison by WAIC and DIC

To compare the models that have been fitted, I will mainly regard the widely applicable information criterion (WAIC) but will also present the deviance information criterion (DIC). DIC is, as the sample size grows, an asymptotic approximation and is widely used but is known to have some problems connected to the fact that it is not fully Bayesian (Vehtari, A, et.al., 2017, p.1414). WAIC however is fully bayesian, which is one of the reasons why WAIC is viewed as an improvement of DIC (Vehtari, A, et.al., 2017, p.1414). WAIC is averaging over the posterior distribution instead of conditioning on a point estimate and uses the entire posterior distribution to calculate its deviance and penalty terms. Hence, a smaller value indicates the better model, which is also the case for DIC (Vehtari, A, et.al., 2017, p.1414).

2.3 Parameter estimation

To estimate the parameters of this model, I will apply the process called Bayesian inference using Integrated Nested Laplace approximation (INLA). INLA is an algorithm created to estimate parameters of LGM models and is a faster alternative to Markov chain Monte Carlo method (MCMC) when regarding spatial and temporal models with large data sets (Martins.G.T, et al., 2013). It is a statistical analyses applied to observed data where the prior distribution needs to be specified for the purpose of obtaining the posterior distribution of the unknown parameter (Blangiardo and Cameletti, 2015, p.58).

Next follows some general theory of the most vital concepts of this method where the aim is to approximate the posterior marginal distributions.

2.3.1 Bayesian inference using INLA

Being related through Bayes theorem, consider the posterior marginal distributions,

(9) $p(\psi | y) = \int p(\psi,\ y | x)\ dx \propto \int p(y|x, \psi)\ p(x|y)\ dx$,

(10) $p(x| y) = \int p(x, \psi | y)\ d\psi = \int p(x|y, \psi) p(\psi|y)\ d\psi$,

where y is the observations, x the latent field, ψ the hyperparameter, $p(\psi| y)$ and $p(x| y)$ are the posterior marginal distributions, $p(y|x, \psi)$ is the likelihood for observation and $p(x, \psi| y)$ is the posterior (Tufvesson, O. 2017, p.13).

The method makes use of Laplace approximation where a density function f(x) of a random variable, with a certain distribution, is regarded as log f(x) by means of a second order Taylor expansion (Blangiardo and Cameletti, 2015, p.105). By solving

(11) $\frac{d\ log\ f(x)}{dx} = 0$, obtaining the mode x^* and

(12) $-1/\ \frac{d^2 logf(x)}{d^2 x}$ for x^* , obtaining σ^{2^*} ,

the Laplace approximation of the given distribution is approximately N(x^* , σ^{2^*}) (Blangiardo and Cameletti, 2015, p.105).

The aim with INLA is to approximate the posterior marginals,

(13) $p(x_j| y) = \int p(x_j,\ \psi| y)\ d\psi = \int p(\psi| y)\ p(x_j| \psi, y)\ d\psi$ and

(14) $p(\psi_k| y) = \int p(\psi| y)\ d\psi_{-k}$

As they are related, it follows from these expressions that approximations needs to be made for $p(\psi| y)$ and $p(x_j| \psi, y)$ (Rue et al., 2009, p.6).

By Laplace approximation is

(15) $p(\psi| y) = \frac{p(x,\psi|y)}{p(x|\psi,y)}$

$$
\begin{aligned}
&= \frac{p(y|x,\psi)p(x,\psi)}{p(y)} \frac{1}{p(x|\psi,y)} \\
&= \frac{p(y|x)p(x|\psi)p(\psi)}{p(y)} \frac{1}{p(x|\psi,y)} \\
&\propto \frac{p(\psi)p(x|\psi)p(y|x)}{p(x|\psi,y)} \\
&\approx \frac{p(\psi)p(x|\psi)p(y|x)}{p^*(x|\psi,y)}, \text{ x} = x^* \quad \text{(the mode)} \\
&= p^*(\psi|\, y) \qquad \text{(the Laplace approximation of } p(\psi|\, y))
\end{aligned}
$$

To approximate $p(x_j|\, \psi, y)$, consider the full density of $x \,|\, y, \psi$, which is approximated to be

$$
\begin{aligned}
(16)\ p(x\,|y,\psi) &\propto p(x\,, y, \psi) \\
&= p(y|x,\psi)\ p(x\,|\psi)\ p(\psi) \\
&\propto p(y|x,\psi)\ p(x\,|\psi), \text{ as a function of x.}
\end{aligned}
$$

From this it follows that the Laplace approximation is

$$(17)\ p(x\,|y,\psi) \approx p^*(x\,|y,\psi) = \mathrm{N}\left(x = x^*,\ -\left[\frac{\partial^2 log p(x^*\,|\psi,y)}{\partial\, x^2}\right]^{-1}\right)$$

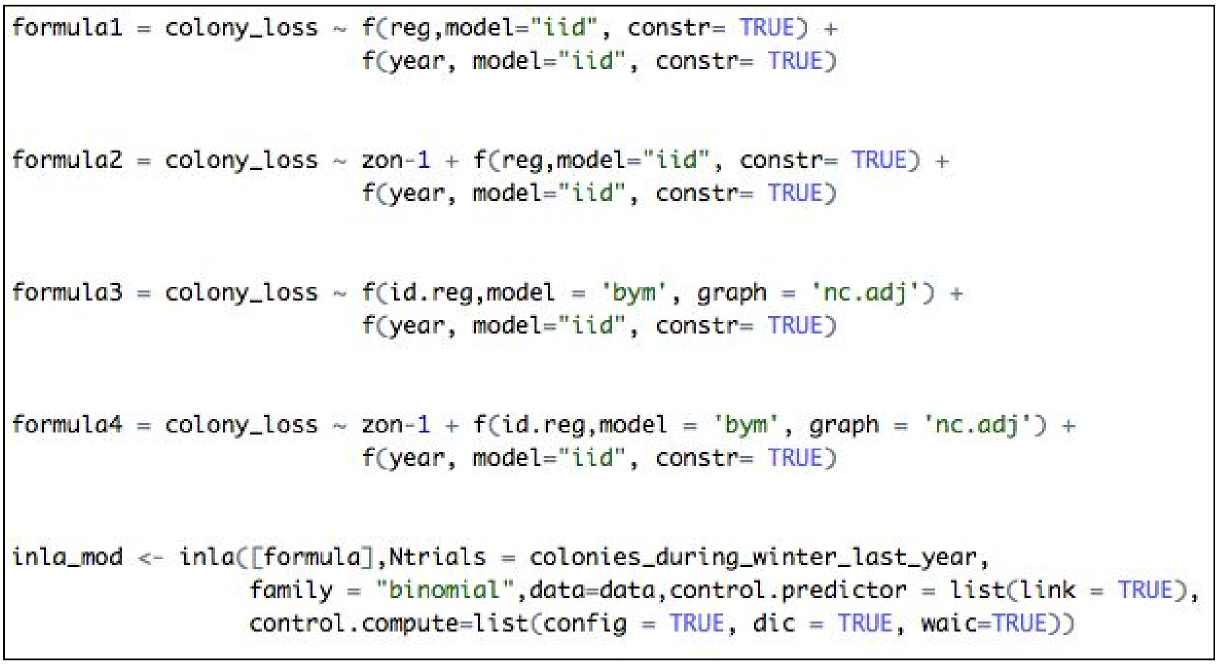
It is possible to do numerical integration of $p^*(x\,|y,\psi)$ to directly derive an approximation of $p(x_j|\, y,\psi)$ which may be a fast method but not always accurate (Rue et al., 2009, p.11). Instead I will be using so called simplified Laplace approximation which is commonly used and it runs by default in the INLA algorithm in R. It is based on a Taylor expansion up to the third order of $p(x\,|\psi,y)$ (Rue et al., 2009, p.12).

With the approximations of $p(x_j|\, y,\psi)$ and $p(\psi|\, y)$, the marginal posteriors $p(x_j|\, y)$ and $p(\psi_k|\, y)$ are obtained through equation (13) and (14) respectively. The integrals are then solved numerically through a series of steps involving finite weighted sums but these are omitted here. (Tufvesson, O. 2017, p.15)

2.4 Implementation

To implement this method I have chosen to work in the software R which is useful for statistical computation and visualisation. The built in package INLA will be used since it is widely used for spatio-temporal models.

Next follows a copy of the implementation of the four models in R, seen as formula 1- 4, and lastly the INLA function that each formula will be put into. v_i and δ_t, seen in model 1 are modeled as i.i.d with a constraint set to zero. In model 3 and 4, $v_i + u_i$ is represented by a bym model which is defined as the i.i.d model + the besag model. z_i is represented by zon.



```
formula1 = colony_loss ~ f(reg,model="iid", constr= TRUE) +
                         f(year, model="iid", constr= TRUE)


formula2 = colony_loss ~ zon-1 + f(reg,model="iid", constr= TRUE) +
                         f(year, model="iid", constr= TRUE)


formula3 = colony_loss ~ f(id.reg,model = 'bym', graph = 'nc.adj') +
                         f(year, model="iid", constr= TRUE)


formula4 = colony_loss ~ zon-1 + f(id.reg,model = 'bym', graph = 'nc.adj') +
                         f(year, model="iid", constr= TRUE)


inla_mod <- inla([formula],Ntrials = colonies_during_winter_last_year,
                family = "binomial",data=data,control.predictor = list(link = TRUE),
                control.compute=list(config = TRUE, dic = TRUE, waic=TRUE))
```

Figure 5. The implementation of the four models and the INLA function in R.

3. Results

The subheadings 3.1 to 3.4 are the results of the INLA approximation i.e the posterior estimates, for model 1 to model 4.

3.1 Model 1

By computing the inverse logit, b_0 results in an mean loss rate of 0.153, for all counties. By looking at $1/\tau_v$ and $1/\tau_\delta$, it can be noted that the variance is bigger for the unstructured spatial effect then it is for the temporal. It means that the loss rate varies more between regions then between years.

Table 2. Posterior estimates (mean, standard deviation (SD) and quantiles) for model 1.

Parameter	Mean	SD	2.5%	50%	97.5%

b_0	-1.711	0.011	-1.734	-1.711	-1.689
$1/\sigma_v^2 = \tau_v$	19.141	6.418	9.216	18.287	34.158
$1/\sigma_\delta^2 = \tau_\delta$	216.385	163.099	34.145	175.669	640.102

3.2 Model 2

By computing the inverse logit, for z_1 - z_5, the resulting mean loss rate for the 5 zones are 0.168, 0.166, 0.123, 0.17 and 0.124 respectively. Similarly to the result of model 1, the variance is bigger for the unstructured spatial effect then it is for the temporal.

Table 3. Posterior estimates (mean, standard deviation (SD) and quantiles) for model 2.

Parameter	Mean	SD	2.5%	50%	97.5%
z_1	-1.598	0.087	-1.771	-1.598	-1.425
z_2	-1.612	0.086	-1.783	-1.612	-1.44
z_3	-1.96	0.102	-2.164	-1.96	-1.757
z_4	-1.581	0.068	-1.715	-1.581	-1.447
z_5	-1.954	0.097	-2.147	-1.953	-1.765
$1/\sigma_v^2 = \tau_v$	32.878	12.968	13.953	30.803	64.006
$1/\sigma_\delta^2 = \tau_\delta$	216.077	163.325	34.213	175.155	641.58

3.3 Model 3

By computing the inverse logit, b_0 results in a mean loss rate of 0.1528, for all regions. The results are similar to model 1 and the variance for the structured spatial effect is much close to 0.

Table 4. Posterior estimates (mean, standard deviation (SD) and

quantiles) for model 3.

Parameter	Mean	SD	2.5%	50%	97.5%
b_0	-1.714	0.056	-1.825	-1.713	-1.604
$1/\sigma_v^2 = \tau_v$	18.292	6.304	8.641	17.422	33.108
$1/\sigma_u^2 = \tau_u$	1813.9	1822.9	113.7	1269.3	6643.5
$1/\sigma_\delta^2 = \tau_\delta$	216.966	163.734	34.437	176.022	643.619

3.4 Model 4

By computing the inverse logit, for z_1 - z_5, the resulting mean loss rate for the 5 zones are 0.172, 0.166, 0.124, 0.17, 0.124 respectively. A similar result to model 2, and again a variance for the structured spatial effect much close to 0.

Table 5. Posterior estimates (mean, standard deviation (SD) and quantiles) for model 4.

Parameter	Mean	SD	2.5%	50%	97.5%
z_1	-1.573	0.114	-1.8	-1.573	-1.347
z_2	-1.6112	0.097	-1.805	-1.612	-1.419
z_3	-1.959	0.113	-2.183	-1.959	-1.735
z_4	-1.589	0.080	-1.742	-1.582	-1.421
z_5	-1.9545	0.109	-2.173	-1.954	-1.742
$1/\sigma_v^2 = \tau_v$	31.17	12.65	12.85	29.10	61.67
$1/\sigma_u^2 = \tau_u$	1863.00	1845.25	130.97	1319.22	6711.36
$1/\sigma_\delta^2 = \tau_\delta$	216.69	163.87	34.42	175.61	644.07

3.5 Model comparison

3.5.1 WAIC and DIC

The resulting information criterions WAIC and DIC for the 4 models are respectively very similar to each other which could indicate that neither of model 2 - 4 are a better alternative to model 1 (table 6). It also means that this result gives no real indication of what model is the better and that regional differences can't be linked to cultivation zones. What is also interesting to observe is the fact that the resulting DIC values are not consistent with those of WAIC. Since the values of WAIC and DIC are so similar, I choose to present further results for model 4 because it's more complex and interesting to look at.

Table 6. Information criterions WAIC and DIC with their respective effective parameters.

Model	WAIC	WAIC p.eff	DIC	DIC p.eff
1	1164.421	180.079	957.352	22.272
2	1164.088	179.646	957.495	22.078
3	1164.372	180.058	957.358	22.301
4	1164.203	179.741	957.530	22.135

3.5.2 Regional loss rate

Few differences can be seen for the regional loss rate between the models (figure 6) which also was concluded in table 6. The loss rate varies between 10.2-19.7 % for model 1, 10.0-19.7% for model 2, 10.2-19.7% for model 3 and 10.0-19.7 % for model 4. Spatial variation however is observed as clusters of counties having similar color.

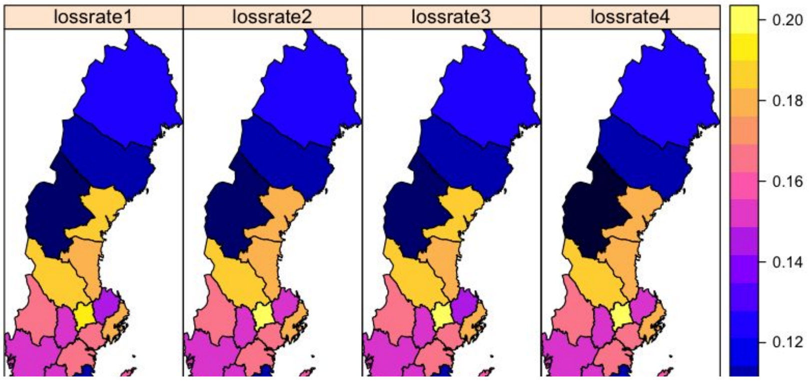


Figure 6. Loss rate due to the four different models. Each map is a representation of the loss rate in the order; model 1 to model 4, from the left.

3.6 Further results

The following results are derived with the estimated parameters of model 4.

3.6.1 Regional loss rate

There is a large variation in mean loss rates for each county (Table 7). There seems to be a spatial pattern across Sweden where clusters of neighbouring counties have similar loss rates ("lossrate4" in figure 6). Uncertainty, seen as 95% probability intervals, in estimates of mean loss rate is varying between counties (figure 7).

Table 7. The mean loss rate of each county and average colonies per county between 2015-2017.

County	Mean	Average colonies	County	Average colonies	Mean	County	Mean	Average colonies
1. Örebro	0.15	1711	8.Norrbotten	501	0.12	15.Östergötland	0.17	3246
2.Jämtland	0.10	305	9.Västerbotten	631	0.11	16.Kalmar	0.11	4307
3.Värmland	0.16	1416	10.Västernorrland	344	0.18	17.Gotland	0.16	1120
4.Kronoberg	0.11	2859	11.Gävleborg	982	0.18	18.Blekinge	0.17	898
5.Jönköping	0.15	3023	12.Uppsala	1945	0.15	19.Skåne	0.19	7310
6.Dalarna	0.19	1094	13.Stockholm	1865	0.18	20.Halland	0.15	2203
7.Västmanland	0.20	664	14.Södermanland	1662	0.17	21.Västra Götaland	0.15	7449

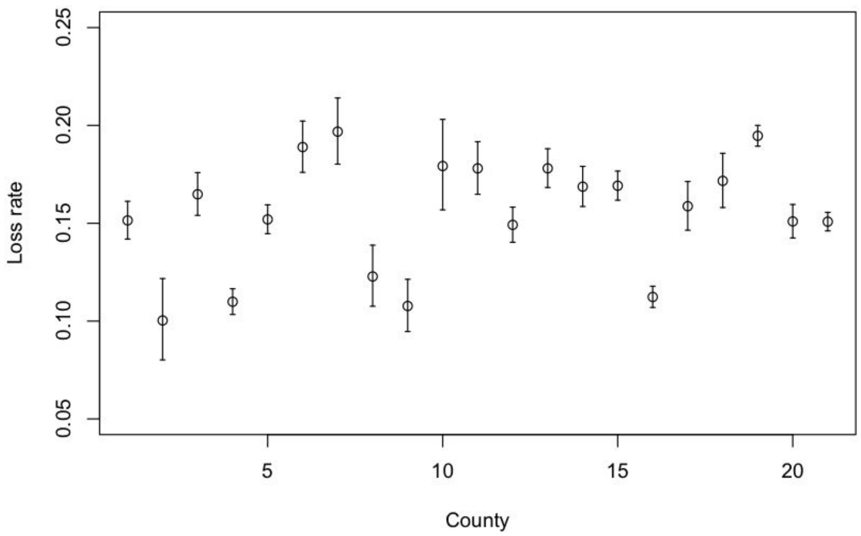


Figure 7. Mean loss rate (circle) with 95% probability intervals for each county, represented by the same order as in table 7.

3.6.2 Zone loss rate

An advantage of the Bayesian model is that estimates of loss rate come as a full probability distribution. The following figures are two ways of presenting the estimated loss rates for the 5 zones. The probability density of mean loss rate on the logit scale for each zone (figure 8, Left panel). Zone 3 and 5 stands out from the others with a slightly smaller mean loss rate. The probability densities can be summarised into 95 % probability intervals, for each zone. The mean loss rate for the 5 zones are 0.172, 0.166, 0.124, 0.17, 0.124 respectively (see paragraph 3.4) with probability intervals (0.14,0.21), (0.14,0.19), (0.10,0.15), (0.15,00.19) and (0.10,0.15) respectively.

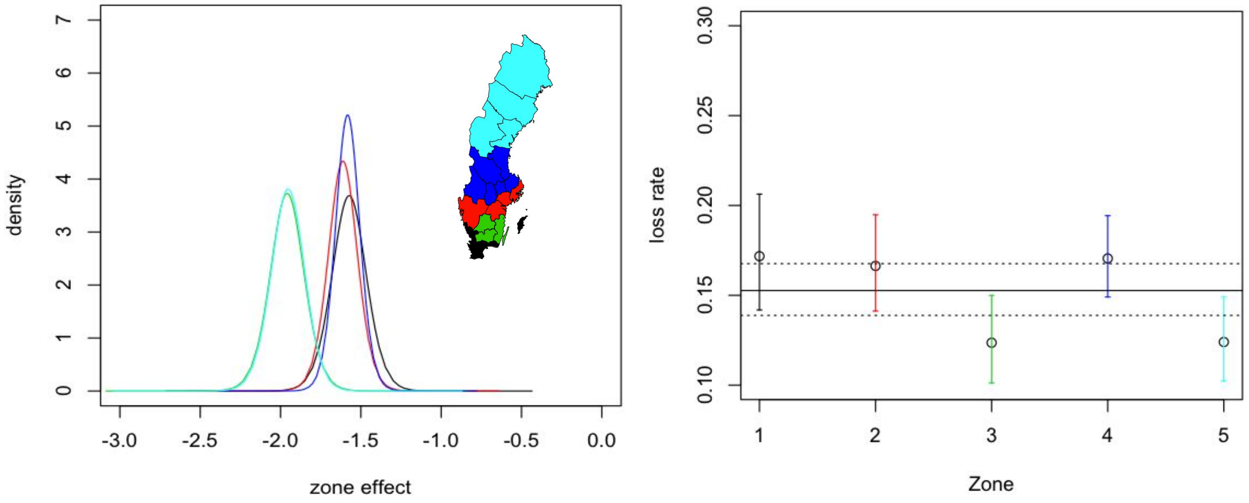


Figure 8. Left panel: Probability density of loss rate on logit scale for each zone.
Right panel: Mean loss rate (circle) with 95% probability intervals for each zone. The black line is the intercept for model 3 with associated confidence interval (dotted lines).

3.6.3 Yearly loss rate

It is possible to observe changes in form of increased loss rate through the 3 years (figure 9). The differences in mean loss rates were increasing from ranges 9.3 to 18.3 % 2015, 9.9% to 19.4 % 2016 and 11.0 to 21.3 % 2017. However, it is not justified to conclude any temporal trends with data from three years only.

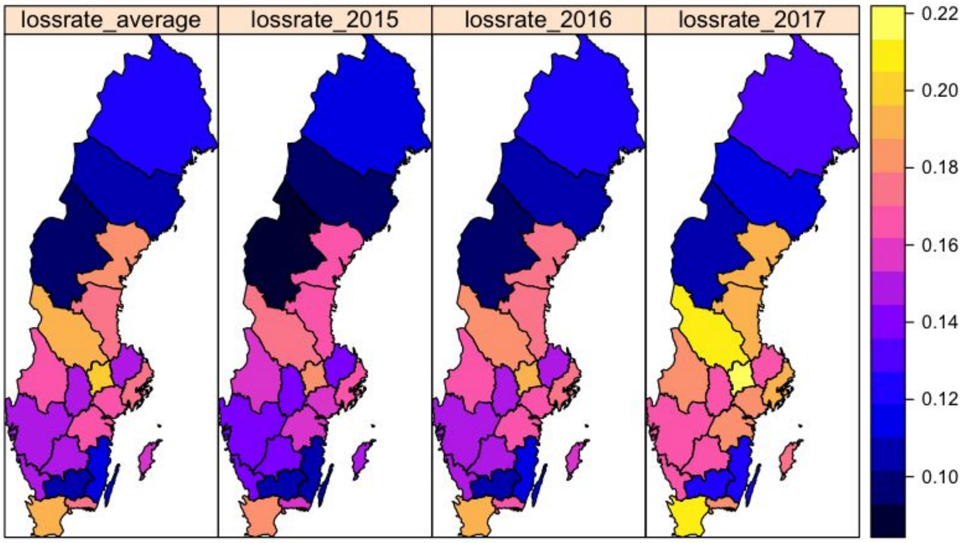


Figure 9. Comparing average loss rate to yearly loss rate (2015-2017).

3.6.4 Historical winter losses

The total winter loss (i.e. not only those that has been reported) is defined as

$$(18) \qquad Y^{*}_{\cdot it} \sim \mathrm{Bin}(m_{it}\,,\, p^{*}_{it})$$

An estimate of the actual number of colonies lost requires an assumption. Under the assumption of an average number of colonies per beekeeper is the same in each region and all members have bees, the total number of colonies in each region *i* and year *t* is defined as

(19) m_{it} = the average number of colonies per reporting members in region *i* year *t**total members in region *i* year *t*.

Historical total winter losses have the mean $m_{it} \times p^{*}_{it}$ i.e the expectation from a binomial distribution. Results are presented and derived with estimations with respect to both zones and counties. The defined total amount of colonies before winter, $\sum_{i=1}^{21} m_{it}$, from 2015-2017, is 270734 colonies and out of these are 27473 estimated to have died during winter.

Table 8. Total colony loss from 2015 to 2017.

Zone	Total colony loss (2015-2017)
1	6237
2	6041
3	4488
4	6192
5	4503
Total	27463

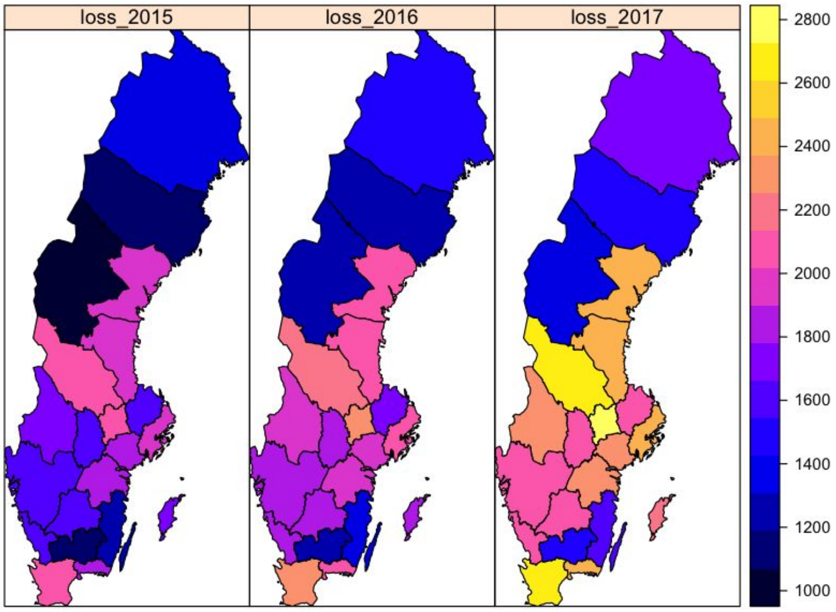


Figure 10. Estimated regional total colony loss per year (2015-2017).

3.6.5 Goodness of fit

A goodness of fit plot shows that the estimates are relatively accurate (figure 11). It plots the expected colony loss for each region i year t, $n_{it} \times p^*_{it}$, against the observed colony loss in each region i year t, where the different colors represents the three years.

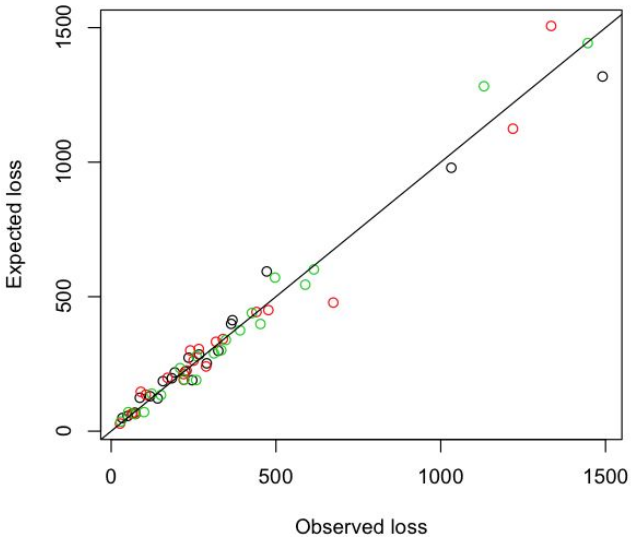


Figure 11. Goodness of fit.

4. Conclusions and discussion

The aim of this thesis was to specify a statistical model for the data on loss rate during winter of honey bees in Sweden, test if regional differences can be linked to the swedish cultivation zones and lastly estimate historical winter losses in Sweden over the last three years.

The result shows that regional differences could not be linked to cultivation zones. Neither did the result show a structured or unstructured spatial effect even though ("lossrate4" in figure 6) suggested it. This could be due to too few observations and that the data is too roughly aggregated and should therefore not be discarded in future studies. The four specified models gave very similar results which then led me to present further results for the more complex model. Unless motivated, the standard procedure would be to choose the simplest of models if no significant effect of new parameters can be established.

Based on the estimated loss rate, the historical total winter loss is assessed to be, that out of 270734 colonies, 27473 died during winter and the mean loss rate is estimated to be between 10.2 and 19.7 %. These results are derived with strong assumptions and should therefore be regarded as a rough upper estimate.

As a result of my assumptions, this thesis is in many ways a simplification of a very complex problem. It was assumed that the colony loss between beekeepers is independent and that all beekeepers in a region has the same probability for winter loss. In reality, there is reasonably a dependency between beekeepers who, for example, have colonies in the same area. It would however, require data on an individual level to estimate variation between beekeepers.

Not only is the average of colonies per county varying a lot (table 7) but it is also established that members holds a varying amount of colonies (figure 12). The illustration itself is an average so there is the possibility of some beekeepers having hundreds of colonies and others few or none. Large versus small scale beekeepers could or do have an effect on the overall colony loss (Van der zee.R, et.al. 2012, p.25). That would mean that the average number I am using could be a too strong of an assumption. However, to estimate its effect also require the data to be collected on an individual level.

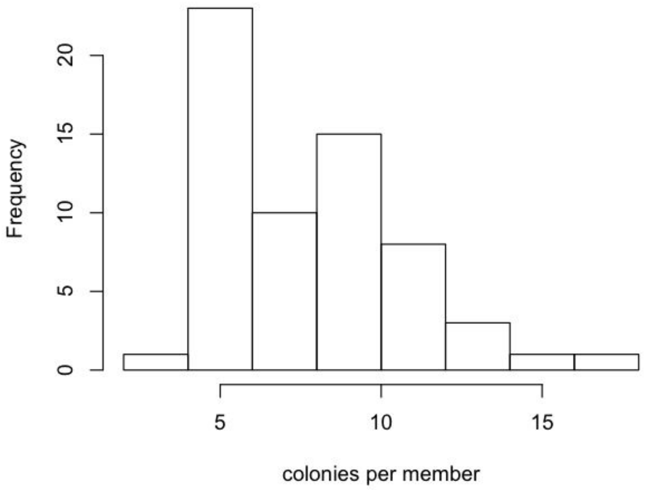


Figure 12. Histogram of the average number of colonies per beekeeper in each region.

The analysis was based on observations from only three years and even though the observed result shows an increasing loss rate, of ranges 9.3 to 18.3 % in 2015, 9.9% to 19.4 % in 2016 and 11 to 21.3 % in 2017, it is not valid to conclude any temporal trends with so few observations. Hence, what the analysis shows when the summary statistics has been collected during a longer period of time, is an interesting aspect of this problem. Apart from being able to find trends, it enables the model to expand, adding regression terms in form of spatial and temporal interactions. Estimates of winter loss rates based on data from several years have stronger properties compared to properties due to changes between regions, especially since the contributing beekeepers may vary a lot from year to year and in each region. Therefore, it would be of interest to do further studies in the future.

5. References

Literature

Blangiardo,M and Cameletti, M. 2015. *Spatial and Spatio-temporal Bayesian Models with R-INLA.* Chichester: John Wiley & sons Ltd.
https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118950203

G. Rodriguez. 2007. *Lecture Notes on Generalized Linear Models.* Princeton university.
http://data.princeton.edu/wws509/notes/c3.pdf

Jordbruksverket, 2009. *Massdöd av bin- samhällsekonomiska konsekvenser och möjliga åtgärder. Rapport 2009:24.*
https://www2.jordbruksverket.se/webdav/files/SJV/trycksaker/Pdf_rapporter/ra09_24.pdf

Martins.G.T, Simpson.D, Lindgren.F and Rue.H. 2013. *Bayesian computing with INLA: new features.* Norwegian University of Science and Technology.
https://www.math.ntnu.no/inla/r-inla.org/papers/CSDAinla_revision.pdf

R Core Team, 2017. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
https://www.R-project.org

Rue, H., Martino, S., and Chopin, N. (2009). *Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations.* Journal of the Royal Statistical Society: Series B (Statistical Methodology), 71(2):319–392.
http://www.statslab.cam.ac.uk/~rjs57/RSS/0708/Rue08.pdf

The R-INLA project. 2018 a. *Independent random noise model.*
https://www.math.ntnu.no/inla/r-inla.org/doc/latent/indep.pd

The R-INLA project. 2018 b. *Besag model for spatial effects.*
https://www.math.ntnu.no/inla/r-inla.org/doc/latent/besag.pdf

Tufvesson, O. 2017. *Spatial statistical modeling of insurance risk - an epidemiologist approach to improved car insurance premiums.* Master's thesis, University of Lund.
http://lup.lub.lu.se/luur/download?func=downloadFile&recordOId=8902318&fileOId=8902319

Van der zee.R, Gray.A, Holzmann.C, Pisa.L, Brodschneider.R, Chlebo.R, Coffey.M.F, Kence.A, Kristiansen.P, Mutinelli.F, Nguyen.B.K, Adjlane.N, Peterson.M, Soroker.V, Topolska.G, Vejsnaes.F, Wilkins.S. 2012. *Standard survey methods for estimating colony losses and explanatory risk factors in Apis mellifera.* In V Dietemann; J D Ellis; P Neumann (Eds) The COLOSS BEEBOOK, Volume I: *Standard methods for Apis mellifera research.* Journal of Apicultural Research 52(4). http://dx.doi.org/10.3896/IBRA.1.52.4.18

Vehtari, A., Gelman, A. & Gabry, J. 2017. *Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC.* Stat Comput (2017) 27: 1413. https://doi-org.ludwig.lub.lu.se/10.1007/s11222-016-9696-4

Pictures

Riksförbundet Svensk trädgård. 2018.
http://www.tradgard.org/svensk_tradgard/zonkarta/zonkarta_stor.html/ (04-06-2018)

Wikimedia Commons, the free media repository. 2018.
https://commons.wikimedia.org/wiki/File:Sveriges_l%C3%A4n_med_namn.jpg (04-06-2018)

Bachelor’s Theses in Mathematical Sciences 2018:K24
ISSN 1654-6229

LUNFMS-4028-2018

Mathematical Statistics
Centre for Mathematical Sciences
Lund University
Box 118, SE-221 00 Lund, Sweden

http://www.maths.lth.se/