*Information Theory Course Project*

# *Filling Data Gaps in Fertilizer Statistics*

*ARASH HATEFI*

*March 2022*

---

## 1. Problem Definition

Swedish official statistics on nitrogen (N) fertilizer use in agriculture are published every year[1]. The public dataset, however, has several missing data which we aim to retrieve from the available values using an information theory analysis. Here, a simplified version of the original dataset is used nonetheless, the same procedure can be applied to the original dataset as well. The dataset, applied method, and results are discussed within this brief report.

## 2. Dataset

A simplified version of the Swedish official statistics on nitrogen (provided by Rasmus Einarsson) was used to demonstrate the procedure. Two types of nitrogen-based fertilizer are presented in the original dataset: *synthetic nitrogen fertilizer* and *animal manure*. Therefore, four different fertilization strategies may exist: (1) only synthetic nitrogen fertilizers, (2) synthetic and manure nitrogen fertilizers in combination, (3) only manure nitrogen fertilizers, or (4) no nitrogen fertilizers. In the simplified dataset, only the two first strategies were considered (synthetic nitrogen fertilizer with or without manure nitrogen fertilizer). The total value of used nitrogen fertilizer is denoted by $m$ (reported in kg). The applications of fertilizers are also grouped by the region ($r$), crop ($c$), and strategy ($s$). As an example, the amount of fertilizer used in region $r$ for crop $c$ with strategy $s$ is denoted by $m_{rcs}$. $m_r$, $m_c$, $m_s$, $m_{rc}$, $m_{rs}$, and $m_{cs}$ are also defined in the same way. We can write

$$m = \sum_{s\in S}\sum_{r\in R}\sum_{c\in C} m_{rcs} \qquad (1)$$

where $S$, $R$, and $C$ are the set of all possible strategies, regions, and crops respectively. Similar equations can be also defined for $m_r$, $m_c$, $m_s$, $m_{rc}$, $m_{rs}$, and $m_{cs}$ as well. The area of crop $c$ harvested in region $r$ and fertilized using strategy $s$ is denoted by $a_{rcs}$ (reported in ha). The values of $m_r$, $m_c$, $m_{rc}$, $m_{cs}$, $m_{rcs}$, and $a_{rcs}$ are defined in separate files, namely `mR.csv`, `mC.csv`, `mRC.csv`, `mCS.csv`, `mRCS.csv`, and `aRCS.csv`. The files `mRCS.csv` and `mRC.csv` have several missing values which we aim to find in this work.

## 3. Data Retreival Methods

Here, the concept of Shannon Entropy from information theory is used for retrieving the missing data. Entropy is a measure of randomness involved in a stochastic process. Higher values of entropy are associated with random distributions with less predictable outcomes. In a Bernoulli process with two outcomes with probabilities $p$ and $p-1$ the entropy is defined as

$$H = -\left[p\log_2 p + (1-p)\log_2(1-p)\right] \qquad (2)$$

taking its maximum value ($s=1$) for $p=0.5$ and minimum value ($s=0$) for $p=0$ or $p=1$ which is expected due to the definition of entropy. Note that in the definition, we assume that $0\log_2 0 \equiv 0$.

---

[1] www.statistikdatabasen.scb.se/pxweb/sv/ssd/START__MI__MI1001

Similarly, for a discrete stochastic process with $n$ possible outcomes $x_1, \dots x_n$ the entropy is

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i) \qquad (3)$$

where $p(x_i)$ is the probability of outcome $x_i$.

Two entropy-based distribution prediction methods namely cross-entropy minimization [1][2] and entropy maximization method [3] were used to tackle the problem. These methods are briefly discussed below.

## 3.1. Entropy Maximization Method

According to this method, the best probability distribution approximating a stochastic process subject to some constraints is the one that maximizes the entropy. Let us assume a stochastic process with $n$ possible outcomes $x_1, \dots x_n$ subject to the following $m$ constraints

$$c_i = f_i[p(x_1), \dots, p(x_n)], \quad i \in \{1, \dots, m\} \quad (4)$$

where $p(x_i)$ is the probability of outcome $x_i$. The best approximations for the values of $p(x_1), \dots, p(x_n)$ are

$$\hat{p}(x_1), \dots, \hat{p}(x_n) = \underset{p(x_1),\dots,p(x_n)}{\operatorname{argmax}} -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i)$$

$$Subject\ to\ \ c_i = 0,\ i \in \{1, \dots, m\} \qquad (5)$$

Therefore, we have a constrained optimization problem which is easy to solve. As an example, using the Lagrange multipliers method we can convert the constrained optimization problem to an unconstrained problem with the following objective function:

$$\mathcal{L}(p(x_1), \dots, p(x_n), \lambda_1, \dots, \lambda_m)$$

$$= \left[-\sum_{i=1}^{n} p(x_i) \log_2 p(x_i)\right] + \sum_{i=0}^{m} \lambda_i c_i \qquad (6)$$

where $\lambda_i$ is the coefficient of the $i$'th constraint. To find the minimum, one can solve the gradient equation:

$$\nabla \mathcal{L} = 0 \qquad (7)$$

## 3.2. Cross-Entropy Minimization Method

The cross-entropy minimization approach provides a model formulation for cases where prior knowledge on the target distribution is available. The method works based on minimizing the Kullback–Leibler divergence (KL divergence) between a prior distribution and a target distribution subjected to several constraints. KL-divergence for two distributions $p_1, p_2, \dots, p_n$ and $q_1, q_2, \dots, q_n$ is defined as

$$KL(p||q) = -\sum_{i=1}^{n} p_i \log_2 \frac{q_i}{p_i} \qquad (8)$$

Which is a measure of how $q$ differs form $p$. KL-divergence is a non-negative quantity and its minimum occurs at zero for two similar distributions. Minimizing the KL divergence between a fixed distribution $p$ and a variable distribution $q$ is equivalent to minimizing their cross-entropy term defined as

$$CE(p, q) = -\sum_{i=1}^{n} p_i \log_2 q_i \qquad (9)$$

Therefore, the method is called cross-entropy minimization.

The complete formulation of the cross-entropy minimization method for estimating a constraint distribution, when a prior distribution $\pi$ is available is

$$\hat{p}(x_1), \dots, \hat{p}(x_n) = \underset{p(x_1),\dots,p(x_n)}{\operatorname{argmin}} -\sum_{i=1}^{n} p(x_i) \log_2 \frac{\pi(x_i)}{p(x_i)}$$

$$Subject\ to \quad c_i = 0,\ i \in \{1, \dots, m\} \qquad (10)$$

which can be also represented as a maximization problem as follow:

$$\hat{p}(x_1), \dots, \hat{p}(x_n) = \underset{p(x_1),\dots,p(x_n)}{argmax} \sum_{i=1}^{n} p(x_i) \log_2 \frac{\pi(x_i)}{p(x_i)}$$

$$Subject\ to \quad c_i = 0,\ i \in \{1, \dots, m\} \qquad (11)$$

Here also the Lagrange multipliers method can be used for finding the distribution as discussed in the previous part.

$$\mathcal{L}(p(x_1), \dots, p(x_n), \lambda_1, \dots \lambda_m) = \left[\sum_{i=1}^{n} P(x_i)\, log_2 \frac{\pi(x_i)}{P(x_i)}\right] + \sum_{i=0}^{m} \lambda_i c_i \quad (12)$$

where $\lambda_i$ is the coefficient of the $i$'th constraint. To find the minimum of the Lagrangean, the gradient equation can be used as discussed in the last section.

## 4. Preparing the Dataset

As mentioned previously, the goal of the project is to estimate the missing $m_{rcs}$ values from other available measures in the dataset. To proceed with an information theory approach, one needs to represent the used fertilizers as a probability distribution. To this end, we divide the values of $m_{rcs}$, $m_r$, $m_c$, $m_s$, $m_{rc}$, $m_{rs}$, and $m_{cs}$ to an estimation of $m$ (total amount of nitrogen fertilizers used) to get $p_{rcs}$, $p_r$, $p_c$, $p_s$, $p_{rc}$, $p_{rs}$, and $p_{cs}$. The estimation for $m$ was obtained using the following formula from the dataset:

$$m \approx \frac{1}{3}\left(\sum_{c\in C} m_c + \sum_{r\in R} m_r + \sum_{c\in C}\sum_{s\in S} m_{cs}\right) \quad (13)$$

The possible constraints to take into account are

$$\sum_{s\in S}\sum_{r\in R} p_{crs} \approx p_c \quad ; \quad \forall c \in C \quad (14)$$

$$\sum_{s\in S}\sum_{c\in C} p_{crs} \approx p_r \quad ; \quad \forall r \in R \quad (15)$$

$$\sum_{r\in R} p_{rcs} \approx p_{cs} \quad ; \quad \forall c \in C, \forall s \in S \quad (16)$$

$$\sum_{s\in S} p_{rcs} \approx p_{cr} \quad ; \quad \forall c \in C, \forall r \in R \quad (17)$$

Note that the approximation sign (≈) in constraints is because of rounding errors that exist in the dataset.

To use the above constraint equations with the Lagrange multipliers method, we represent them in a standard form with the right side being equal to zero. We then store the left sides (that are basically zero quantities) to separate sets:

$$\mathbb{C}_1 = \{\forall c \in C | \textstyle\sum_{s\in S}\sum_{r\in R} p_{crs} - p_c\} \quad (18)$$

$$\mathbb{C}_2 = \{\forall r \in R | \textstyle\sum_{s\in S}\sum_{c\in C} p_{crs} - p_r\} \quad (19)$$

$$\mathbb{C}_3 = \{\forall c \in C, \forall s \in S | \textstyle\sum_{r\in R} p_{rcs} - p_{cs}\} \quad (20)$$

$$\mathbb{C}_4 = \{\forall c \in C, \forall r \in R | \textstyle\sum_{s\in S} p_{rcs} - p_{cr}\} \quad (21)$$

Also, as some $m_{rcs}$ values are available in the dataset, one can also add the following set of constraints to the problem:

$$[p_{crs}]_i = p(x_i) \quad ; \quad i \in I \quad (22)$$

Where $[p_{crs}]_i$ is the value of $p_{crs}$ in the $i$th row and $I$ is the set of row indices that their values are available. The above set of constraints can be written as

$$\mathbb{C}_5 = \{i \in I | [p_{crs}]_i - p(x_i)\} \quad (23)$$

In the case of the used dataset, the total number of independent constraints is larger than the number of variables, making the problem over-constraint. This is because of the error in the dataset so that, as an example, the sum of $m_{rc}$ values for a specific region $r^*$ is not exactly equal to $m_{r^*}$, but an approximation to it. There are two approaches for handling the issue. First, one can simply ignore some of the constraints (say $\mathbb{C}_1$ and $\mathbb{C}_2$ as they are to some extent incorporated in $\mathbb{C}_3$ and $\mathbb{C}_4$) and use a subset of independent constraints. Another approach is to directly solve the over-constraint problem and accept a certain amount of deviation from each of the constraints. The second approach is expected to perform better as all the constraints are approximations and the more constraint we use, the lower would be the effect of the noise in the dataset (like rounding errors) on the solution.

Here, we solve the problem using four different sets of constraints as shown below:

$$\mathbb{C}_I = \mathbb{C}_1 \cup \mathbb{C}_2 \quad (24)$$

$$\mathbb{C}_{II} = \mathbb{C}_3 \cup \mathbb{C}_4 \quad (25)$$

$$\mathbb{C}_{III} = \mathbb{C}_1 \cup \mathbb{C}_2 \cup \mathbb{C}_3 \cup \mathbb{C}_4 \quad (26)$$

$$\mathbb{C}_{IV} = \mathbb{C}_1 \cup \mathbb{C}_2 \cup \mathbb{C}_3 \cup \mathbb{C}_4 \cup \mathbb{C}_5 \quad (27)$$

# 5. Finding the Missing values

The missing values are found using the entropy maximization and cross-entropy minimization methods, each applied to problems with constraints set $\mathbb{C}_I$, $\mathbb{C}_{II}$, $\mathbb{C}_{III}$ or $\mathbb{C}_{IV}$. For the cross-entropy minimization method we also need a prior distribution. We solve the problem with two different priors. In the first attempt we consider a uniform distribution, as the distribution with the maximum uncertainty for the prior knowledge. For a more advanced prior we assure the amount of nitrogen fertilizer is proportional to the harvesting area.

$$m_{crs} \propto a_{crs} \quad (28)$$

Therefore, one can find the prior distribution as

$$\pi_{crs} = \frac{a_{crs}}{\sum_{c\in C}\sum_{s\in S}\sum_{r\in R} a_{crs}} \quad (29)$$

For solving each of the constraint optimization problems (including the over-constraint ones), we minimized the second norm of the Lagrangean gradient. The new objective is

$$\hat{p}(x_1), \dots, \hat{p}(x_n) = \underset{p(x_1),\dots,p(x_n)}{\operatorname{argmin}} \|\nabla \mathcal{L}\| \quad (30)$$

Which is an unconstraint non-linear optimization problem easily solvable by Matlab. The set of parameters used in each optimization process is listed in table 1.

***Table 1: proposed settings for solving the problem***

| *Name* | *Method* | *Constraints* | *Prior* |
|---|---|---|---|
| *Setting 1* | *Entropy Maximization* | $\mathbb{C}_I$ | - |
| *Setting 2* | *Entropy Maximization* | $\mathbb{C}_{II}$ | - |
| *Setting 3* | *Entropy Maximization* | $\mathbb{C}_{III}$ | - |
| *Setting 4* | *Entropy Maximization* | $\mathbb{C}_{IV}$ | - |
| *Setting 5* | *Cross-Entropy Minimization* | $\mathbb{C}_I$ | *Uniform* |
| *Setting 6* | *Cross-Entropy Minimization* | $\mathbb{C}_I$ | $m_{crs} \propto a_{crs}$ |
| *Setting 7* | *Cross-Entropy Minimization* | $\mathbb{C}_{II}$ | *Uniform* |
| *Setting 8* | *Cross-Entropy Minimization* | $\mathbb{C}_{II}$ | $m_{crs} \propto a_{crs}$ |
| *Setting 9* | *Cross-Entropy Minimization* | $\mathbb{C}_{III}$ | *Uniform* |
| *Setting 10* | *Cross-Entropy Minimization* | $\mathbb{C}_{III}$ | $m_{crs} \propto a_{crs}$ |
| *Setting 11* | *Cross-Entropy Minimization* | $\mathbb{C}_{IV}$ | *Uniform* |
| *Setting 12* | *Cross-Entropy Minimization* | $\mathbb{C}_{IV}$ | $m_{crs} \propto a_{crs}$ |

Also, the predicted values using the entropy maximization and cross-entropy minimization methods are available in tables 2 and 3 respectively.

***Table 2: Found Values using the Entropy Maximization Method***

| *Region* | *Crop* | *Strategy*[*] | *M* | | | | |
|---|---|---|---|---|---|---|---|
| | | | *Setting 1* | *Setting 2* | *Setting 3* | *Setting 4* | *Real Values* |
| *PO1* | *Cereals* | *S* | 9745 | 17063 | 17127 | 18163 | 18248 |
| *PO1* | *Cereals* | *S & M* | 9745 | 3753 | 3806 | 2894 | 2907 |
| *PO1* | *Ley Silage* | *S* | 4801 | 1159 | 1223 | 1225 | 1215 |
| *PO1* | *Ley Silage* | *S & M* | 4801 | 2468 | 2591 | 2444 | 2440 |
| *PO1* | *Ley Pasture* | *S* | 231 | 203 | 194 | 117 | - |
| *PO1* | *Ley Pasture* | *S & M* | 231 | 164 | 168 | 50 | - |
| *PO1* | *Other* | *S* | 2977 | 7444 | 7638 | 8708 | 8776 |

[*] S: synthetic, S & M: synthetic and manure

***Table 3 (Continue)***

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| *PO1* | *Other* | *S & M* | 2977 | 2611 | 2652 | 1767 | 1778 |
| *PO2* | *Cereals* | *S* | 6611 | 8285 | 8397 | 7356 | 7376 |
| *PO2* | *Cereals* | *S & M* | 6611 | 1822 | 1866 | 2985 | 2956 |
| *PO2* | *Ley Silage* | *S* | 3257 | 2435 | 2522 | 1980 | 1972 |
| *PO2* | *Ley Silage* | *S & M* | 3257 | 5183 | 5343 | 5786 | 5775 |
| *PO2* | *Ley Pasture* | *S* | 157 | 260 | 151 | 229 | 256 |
| *PO2* | *Ley Pasture* | *S & M* | 157 | 210 | 130 | 24 | - |
| *PO2* | *Other* | *S* | 2020 | 3929 | 4100 | 2814 | 2840 |
| *PO2* | *Other* | *S & M* | 2020 | 1378 | 1423 | 2877 | 2876 |
| *PO3* | *Cereals* | *S* | 8911 | 18740 | 18827 | 19657 | 19685 |
| *PO3* | *Cereals* | *S & M* | 8911 | 4122 | 4184 | 3478 | 3436 |
| *PO3* | *Ley Silage* | *S* | 4390 | 1593 | 1643 | 1907 | 1887 |
| *PO3* | *Ley Silage* | *S & M* | 4390 | 3391 | 3481 | 3160 | 3146 |
| *PO3* | *Ley Pasture* | *S* | 212 | 195 | 196 | 175 | 218 |
| *PO3* | *Ley Pasture* | *S & M* | 212 | 158 | 170 | 28 | - |
| *PO3* | *Other* | *S* | 2722 | 2727 | 2847 | 3334 | 3364 |
| *PO3* | *Other* | *S & M* | 2722 | 956 | 988 | 654 | 662 |
| *PO4* | *Cereals* | *S* | 8414 | 17405 | 17501 | 18673 | 18701 |
| *PO4* | *Cereals* | *S & M* | 8414 | 3828 | 3890 | 2905 | 2873 |
| *PO4* | *Ley Silage* | *S* | 4145 | 1521 | 1577 | 3017 | 2992 |
| *PO4* | *Ley Silage* | *S & M* | 4145 | 3238 | 3341 | 1842 | 1830 |
| *PO4* | *Ley Pasture* | *S* | 200 | 187 | 190 | 225 | 269 |
| *PO4* | *Ley Pasture* | *S & M* | 200 | 151 | 164 | 13 | - |
| *PO4* | *Other* | *S* | 2571 | 2739 | 2859 | 3495 | 3518 |
| *PO4* | *Other* | *S & M* | 2571 | 960 | 992 | 505 | 513 |
| *PO5* | *Cereals* | *S* | 5546 | 4427 | 4590 | 3699 | 3726 |
| *PO5* | *Cereals* | *S & M* | 5546 | 973 | 1020 | 2013 | 1934 |
| *PO5* | *Ley Silage* | *S* | 2732 | 4062 | 4183 | 2476 | 2450 |
| *PO5* | *Ley Silage* | *S & M* | 2732 | 8645 | 8859 | 10424 | 10400 |
| *PO5* | *Ley Pasture* | *S* | 132 | 314 | 59 | 497 | 490 |
| *PO5* | *Ley Pasture* | *S & M* | 132 | 254 | 51 | 61 | - |
| *PO5* | *Other* | *S* | 1694 | 750 | 919 | 570 | 594 |
| *PO5* | *Other* | *S & M* | 1694 | 263 | 320 | 472 | 476 |
| *PO6* | *Cereals* | *S* | 1512 | 2155 | 1802 | 1402 | 1436 |
| *PO6* | *Cereals* | *S & M* | 1512 | 474 | 400 | 933 | 931 |
| *PO6* | *Ley Silage* | *S* | 745 | 833 | 825 | 1323 | 1312 |
| *PO6* | *Ley Silage* | *S & M* | 745 | 1774 | 1748 | 1232 | - |
| *PO6* | *Ley Pasture* | *S* | 36 | 168 | 180 | 233 | - |
| *PO6* | *Ley Pasture* | *S & M* | 36 | 136 | 156 | 99 | - |
| *PO6* | *Other* | *S* | 462 | 637 | 359 | 284 | - |
| *PO6* | *Other* | *S & M* | 462 | 223 | 125 | 96 | - |
| *PO7* | *Cereals* | *S* | 891 | 737 | 720 | 343 | - |
| *PO7* | *Cereals* | *S & M* | 891 | 162 | 160 | 148 | - |
| *PO7* | *Ley Silage* | *S* | 439 | 772 | 677 | 592 | - |
| *PO7* | *Ley Silage* | *S & M* | 439 | 1644 | 1434 | 1757 | - |
| *PO7* | *Ley Pasture* | *S* | 21 | 168 | 157 | 168 | - |
| *PO7* | *Ley Pasture* | *S & M* | 21 | 136 | 136 | 71 | - |
| *PO7* | *Other* | *S* | 272 | 637 | 313 | 203 | - |
| *PO7* | *Other* | *S & M* | 272 | 223 | 109 | 69 | - |
| *PO8* | *Cereals* | *S* | 993 | 757 | 721 | 426 | - |
| *PO8* | *Cereals* | *S & M* | 993 | 167 | 160 | 183 | - |
| *PO8* | *Ley Silage* | *S* | 489 | 914 | 772 | 970 | - |
| *PO8* | *Ley Silage* | *S & M* | 489 | 1946 | 1634 | 1840 | 1826 |
| *PO8* | *Ley Pasture* | *S* | 24 | 168 | 195 | 98 | - |
| *PO8* | *Ley Pasture* | *S & M* | 24 | 136 | 169 | 42 | - |
| *PO8* | *Other* | *S* | 303 | 637 | 389 | 120 | - |
| *PO8* | *Other* | *S & M* | 303 | 223 | 135 | 40 | - |

***Table 3: Found Values using the Cross-Entropy Minimization Method***

| Region | Crop | Strategy* | M | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Setting 5 | Setting 6 | Setting 7 | Setting 8 | Setting 9 | Setting 10 | Setting 11 | Setting 12 | Real Values |
| PO1 | Cereals | S | 9724 | 16917 | 14575 | 17966 | 16845 | 17956 | 17411 | 18221 | 18248 |
| PO1 | Cereals | S & M | 9724 | 3983 | 3317 | 3046 | 3761 | 3067 | 2953 | 2869 | 2907 |
| PO1 | Ley Silage | S | 4786 | 1372 | 1156 | 1258 | 1262 | 1272 | 1197 | 1201 | 1215 |
| PO1 | Ley Silage | S & M | 4786 | 2324 | 2356 | 2376 | 2670 | 2396 | 2541 | 2448 | 2440 |
| PO1 | Ley Pasture | S | 269 | 227 | 42 | 200 | 304 | 187 | 819 | 183 | - |
| PO1 | Ley Pasture | S & M | 269 | 70 | 39 | 69 | 255 | 76 | 384 | 73 | - |
| PO1 | Other | S | 2965 | 8315 | 6013 | 8647 | 7466 | 8634 | 8525 | 8759 | 8776 |
| PO1 | Other | S & M | 2965 | 2301 | 2553 | 1828 | 2641 | 1832 | 1514 | 1765 | 1778 |
| PO2 | Cereals | S | 6595 | 6586 | 7711 | 7393 | 8412 | 7413 | 7193 | 7374 | 7376 |
| PO2 | Cereals | S & M | 6595 | 3520 | 1747 | 2845 | 1877 | 2874 | 2923 | 2945 | 2956 |
| PO2 | Ley Silage | S | 3246 | 2037 | 2283 | 1795 | 2546 | 1810 | 2068 | 1958 | 1972 |
| PO2 | Ley Silage | S & M | 3246 | 6022 | 4644 | 5922 | 5382 | 5953 | 5775 | 5790 | 5775 |
| PO2 | Ley Pasture | S | 182 | 244 | 271 | 274 | 45 | 267 | 204 | 225 | 256 |
| PO2 | Ley Pasture | S & M | 182 | 49 | 252 | 61 | 37 | 70 | 181 | 66 | - |
| PO2 | Other | S | 2011 | 2410 | 3919 | 2798 | 4006 | 2803 | 2800 | 2842 | 2840 |
| PO2 | Other | S & M | 2011 | 3220 | 1662 | 2857 | 1417 | 2871 | 2648 | 2877 | 2876 |
| PO3 | Cereals | S | 8891 | 18351 | 15922 | 19476 | 18484 | 19483 | 18932 | 19700 | 19685 |
| PO3 | Cereals | S & M | 8891 | 4704 | 3626 | 3596 | 4127 | 3623 | 3629 | 3441 | 3436 |
| PO3 | Ley Silage | S | 4376 | 2087 | 1542 | 1946 | 1707 | 1963 | 2031 | 1890 | 1887 |
| PO3 | Ley Silage | S & M | 4376 | 2971 | 3142 | 3090 | 3610 | 3108 | 3223 | 3173 | 3146 |
| PO3 | Ley Pasture | S | 246 | 226 | 120 | 230 | 340 | 225 | 308 | 184 | 218 |
| PO3 | Ley Pasture | S & M | 246 | 34 | 112 | 39 | 285 | 45 | 170 | 41 | - |
| PO3 | Other | S | 2711 | 3280 | 2522 | 3328 | 2687 | 3338 | 2978 | 3363 | 3364 |
| PO3 | Other | S & M | 2711 | 814 | 1066 | 631 | 950 | 635 | 728 | 656 | 662 |
| PO4 | Cereals | S | 8395 | 17686 | 14848 | 18806 | 17206 | 18844 | 17971 | 18680 | 18701 |
| PO4 | Cereals | S & M | 8395 | 3432 | 3379 | 2628 | 3841 | 2652 | 3020 | 2842 | 2873 |
| PO4 | Ley Silage | S | 4132 | 3505 | 1458 | 3004 | 1637 | 3038 | 2953 | 2984 | 2992 |
| PO4 | Ley Silage | S & M | 4132 | 1887 | 2971 | 1804 | 3461 | 1819 | 2084 | 1848 | 1830 |
| PO4 | Ley Pasture | S | 232 | 242 | 34 | 295 | 299 | 283 | 176 | 233 | 269 |
| PO4 | Ley Pasture | S & M | 232 | 23 | 32 | 31 | 250 | 35 | 188 | 37 | - |
| PO4 | Other | S | 2560 | 3319 | 2546 | 3522 | 2708 | 3529 | 3135 | 3517 | 3518 |
| PO4 | Other | S & M | 2560 | 565 | 1077 | 458 | 958 | 460 | 696 | 508 | 513 |
| PO5 | Cereals | S | 5532 | 3569 | 4371 | 3625 | 4525 | 3675 | 3667 | 3720 | 3726 |
| PO5 | Cereals | S & M | 5532 | 2646 | 987 | 1936 | 1009 | 1976 | 1876 | 1925 | 1934 |
| PO5 | Ley Silage | S | 2722 | 2398 | 3749 | 2317 | 4114 | 2334 | 2530 | 2454 | 2450 |
| PO5 | Ley Silage | S & M | 2722 | 9788 | 7614 | 10550 | 8696 | 10596 | 10257 | 10431 | 10400 |
| PO5 | Ley Pasture | S | 153 | 486 | 251 | 526 | 334 | 412 | 413 | 465 | 490 |
| PO5 | Ley Pasture | S & M | 153 | 137 | 233 | 164 | 280 | 151 | 170 | 145 | - |
| PO5 | Other | S | 1686 | 527 | 715 | 511 | 816 | 533 | 681 | 566 | 594 |
| PO5 | Other | S & M | 1686 | 658 | 303 | 488 | 289 | 511 | 435 | 487 | 476 |
| PO6 | Cereals | S | 1526 | 1348 | 2989 | 1482 | 1989 | 1476 | 1536 | 1432 | 1436 |
| PO6 | Cereals | S & M | 1526 | 1036 | 674 | 820 | 444 | 823 | 731 | 918 | 931 |
| PO6 | Ley Silage | S | 751 | 1416 | 958 | 1330 | 788 | 1324 | 1582 | 1282 | 1312 |
| PO6 | Ley Silage | S & M | 751 | 1203 | 1953 | 1261 | 1666 | 1252 | 817 | 1326 | - |
| PO6 | Ley Pasture | S | 42 | 104 | 1671 | 163 | 96 | 158 | 311 | 147 | - |
| PO6 | Ley Pasture | S & M | 42 | 10 | 1553 | 17 | 80 | 20 | 146 | 18 | - |
| PO6 | Other | S | 465 | 298 | 872 | 421 | 623 | 406 | 597 | 307 | - |
| PO6 | Other | S & M | 465 | 94 | 370 | 101 | 220 | 98 | 374 | 94 | - |
| PO7 | Cereals | S | 914 | 368 | 2619 | 433 | 946 | 362 | 1087 | 277 | - |
| PO7 | Cereals | S & M | 914 | 346 | 591 | 293 | 213 | 247 | 171 | 262 | - |

* S: synthetic, S & M: synthetic and manure

***Table 3 (Continue)***

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *PO7* | *Ley Silage* | *S* | 450 | 852 | 922 | 836 | 616 | 817 | 525 | 757 | - |
| *PO7* | *Ley Silage* | *S & M* | 450 | 1428 | 1881 | 1563 | 1303 | 1525 | 1584 | 1639 | - |
| *PO7* | *Ley Pasture* | *S* | 25 | 93 | 1671 | 155 | 85 | 154 | 163 | 142 | - |
| *PO7* | *Ley Pasture* | *S & M* | 25 | 9 | 1553 | 17 | 71 | 20 | 76 | 18 | - |
| *PO7* | *Other* | *S* | 279 | 120 | 872 | 181 | 556 | 178 | 313 | 134 | - |
| *PO7* | *Other* | *S & M* | 279 | 32 | 370 | 37 | 197 | 36 | 196 | 35 | - |
| *PO8* | *Cereals* | *S* | 1005 | 453 | 2638 | 564 | 991 | 516 | 1086 | 369 | - |
| *PO8* | *Cereals* | *S & M* | 1005 | 299 | 595 | 269 | 223 | 248 | 171 | 247 | - |
| *PO8* | *Ley Silage* | *S* | 495 | 965 | 1004 | 945 | 738 | 912 | 512 | 961 | - |
| *PO8* | *Ley Silage* | *S & M* | 495 | 1740 | 2048 | 1902 | 1561 | 1830 | 1901 | 1850 | 1826 |
| *PO8* | *Ley Pasture* | *S* | 28 | 49 | 1671 | 85 | 85 | 81 | 193 | 81 | - |
| *PO8* | *Ley Pasture* | *S & M* | 28 | 22 | 1553 | 43 | 72 | 48 | 90 | 47 | - |
| *PO8* | *Other* | *S* | 306 | 61 | 872 | 96 | 557 | 91 | 371 | 73 | - |
| *PO8* | *Other* | *S & M* | 306 | 30 | 370 | 37 | 197 | 35 | 232 | 36 | - |

# 6. Interpreting the Results

The coefficient of determination ($r^2$) is used as a measure for comparing the accuracy of different solutions. $r^2$ is a measure of success in predicting correct values which takes its maximum value at one when predictions are identical to real values. The mathematical formula for $r^2$ is as follow:

$$r^2 = 1 - \frac{\sum_{c\in C}\sum_{s\in S}\sum_{r\in R}\left(\hat{p}_{crs} - p_{crs}\right)^2}{\sum_{c\in C}\sum_{s\in S}\sum_{r\in R}\left(\bar{p} - p_{crs}\right)^2} \qquad (31)$$

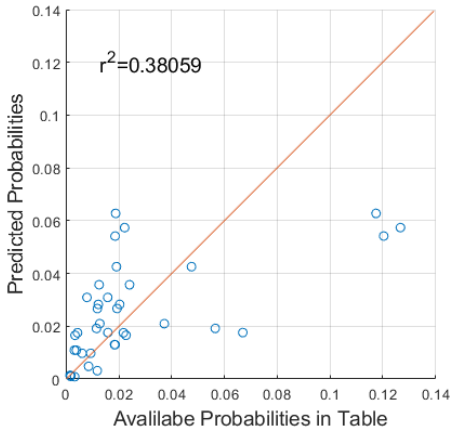
where $\bar{p}$ is the average of $p_{crs}$ values and $\hat{p}_{crs}$ are the predicted values. As some $p_{crs}$ values are missing from the table, only the available ones with their corresponding predictions are considered in predicting $r^2$.

The values of $r^2$ for each of the solutions are shown in table 3.

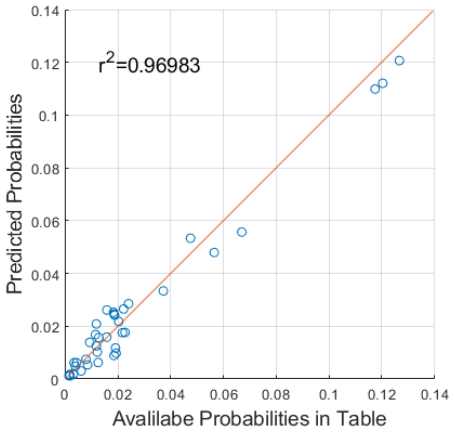
Using the cross-entropy minimization method with uniform distribution, we get approximately the same results from entropy maximization. This is due to the fact that the maximum of the entropy function for a random process occurs at uniform distribution (In other words, by using the uniform prior distribution, equation 12 becomes same as equation 6). The slight deviations are due to the numerical errors. By considering the amount of fertilizer being proportional to the harvesting area, the cross-entropy minimization method outperforms the entropy maximization which is expected because in reality, the amount of fertilizer correlates with the harvesting area. Finally, adding more constraints generally improves the accuracy of prediction and the highest accuracies are achieved in the over-constraint optimizations, where all five at of constraints are taken into account. The drawback of adding the fifth set of constraints ($\mathbb{C}_5$) is that using these extra constraints makes it difficult to estimate the accuracy of predictions for missing values from the accuracy of presented values. This is because both methods have a better performance on predicting the available values which were fed to them within constraints in $\mathbb{C}_5$. The predicted values ($\hat{p}_{crs}$) versus the available ones in the table ($p_{crs}$) for different settings are shown in figure 1.

***Table 4: $r^2$ Values for Different Settings***

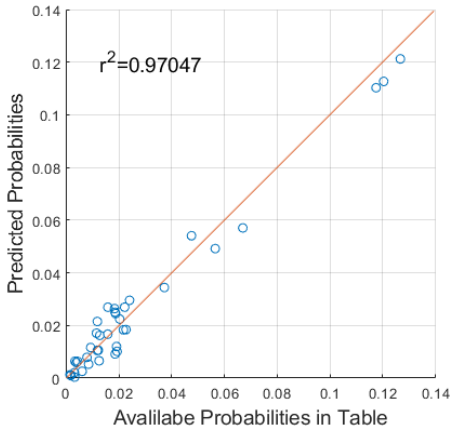
| ***Solution*** | ***$r^2$*** |
|---|---|
| *Setting 1* | *0.38059* |
| *Setting 2* | *0.96983* |
| *Setting 3* | *0.97047* |
| *Setting 4* | *0.99878* |
| *Setting 5* | *0.38055* |
| *Setting 6* | *0.98844* |
| *Setting 7* | *0.94346* |
| *Setting 8* | *0.99955* |
| *Setting 9* | *0.96721* |
| *Setting 10* | *0.99953* |
| *Setting 11* | *0.99711* |
| *Setting 12* | *0.99999* |



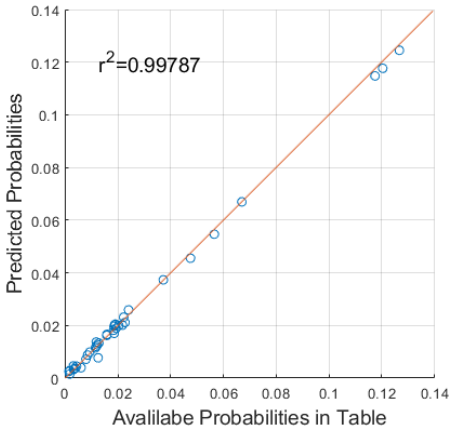

*(a) Setting 1*



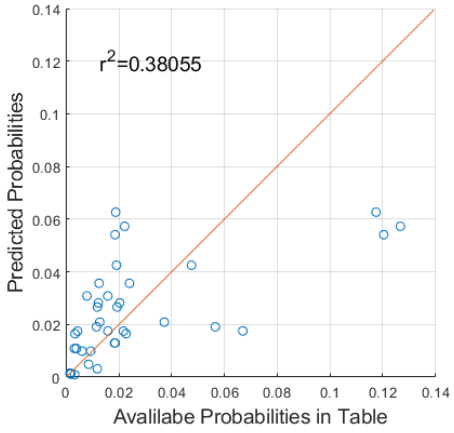

*(b) Setting 2*



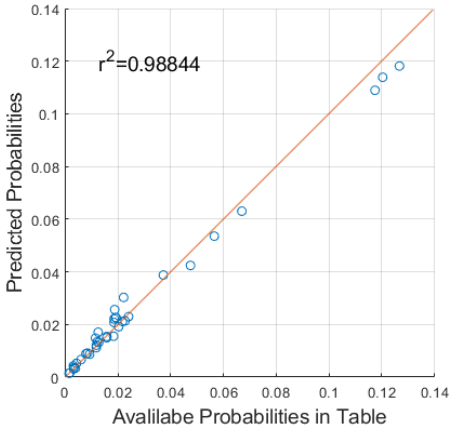

*(c) Setting 3*




*(d) Setting 4*

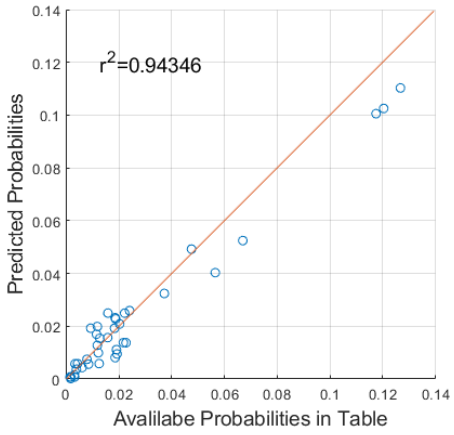



*(a) Setting 5*

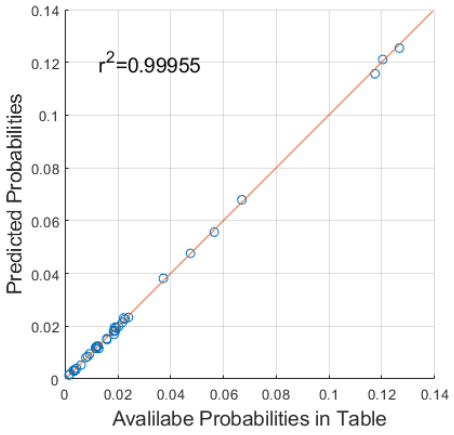



*(b) Setting 6*

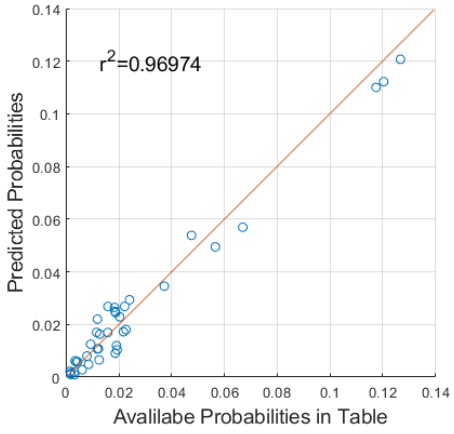
***Figure 1: The predicted values ($\hat{p}_{crs}$) versus the available ones in the table ($p_{crs}$) for different settings***



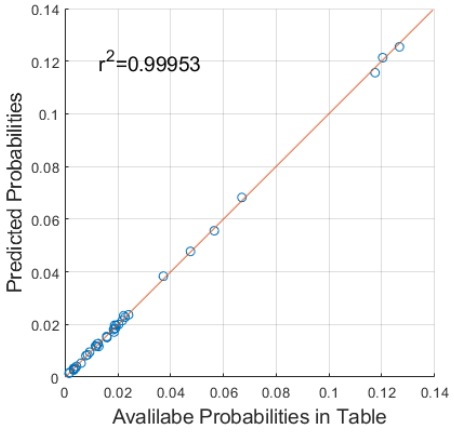

*(c) Setting 7*



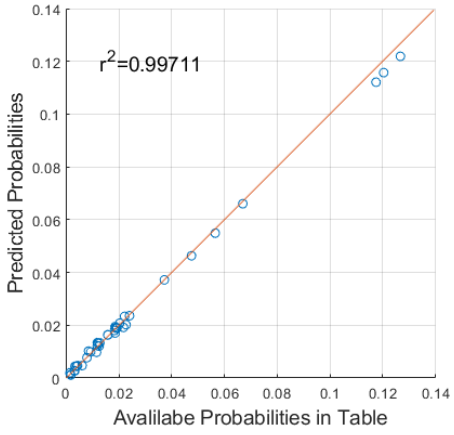

*(d) Setting 8*



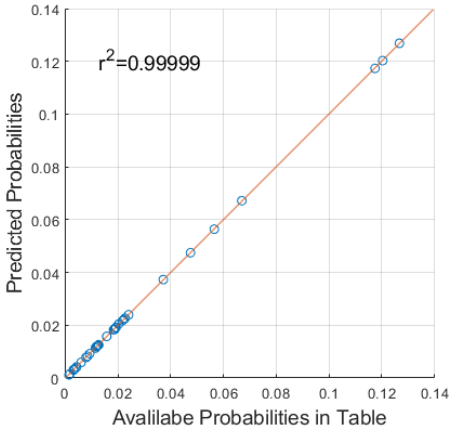

*(e) Setting 9*




*(f) Setting 10*




*(e) Setting 11*




*(f) Setting 12*

***Figure 1 (Continue)***

## 7. Conclusion

In this work, cross-entropy minimization and entropy maximization methods were used for estimating the missing data in the Swedish statistics on nitrogen fertilizers. Both methods did an acceptable job of predicting the missing data, while the first one outperformed the second when used with a realistic prior distribution. It was seen that increasing the number of constraints can be beneficial for reaching higher predictions accuracy, even though it may result in an over-constraint problem which can be relaxed by considering the constraints as approximations to reality.

## References


[1] You, L. and S. Wood (2005). Assessing the Spatial Distribution of Crop Areas Using a Cross-Entropy Method. International Journal of Applied Earth Observation and Geoinformation. Bridging Scales and Epistemologies Linking Local Knowledge with Global Science in Multi-Scale Assessments 7(4), pp. 310–323.

[2] You, L., S. Wood, U. Wood-Sichra, and W. Wu (2014). Generating Global Crop Distribution Maps: From Census to Grid. Agricultural Systems 127, pp. 53–60.

[3] *Guiasu, S. and Shenitzer, A. (1985). The principle of maximum entropy. The Mathematical Intelligencer. pp. 42–48.*