ISSN: 1803-3814

MENDEL Soft Computing

HEURISTIC APPROACH TO MULTIVARIATE INVERSE PREDICTION PROBLEM USING DATA RECONCILIATION

Martin Rosecký[1], Radovan Šomplák[2,3], František Janošt'ák[3], Josef Bednář[1]

Brno University of Technology, Faculty of Mechanical Engineering
[1]Institute of Mathematics
[2]Sustainable Process Integration Laboratory - SPIL, NETME Centre
[3]Institute of Process Engineering
Technická 2896/2, 616 69 Brno
Czech Republic
Martin.Rosecky@vutbr.cz

Abstract: *Some engineering waste management tasks require a complete data sets of its production. However, these sets are not available in most cases. Whether they are not archiving at all or are unavailable for their sensitivity. This article deals with the issue of incomplete datasets at the microregional level. For estimates, the data from higher territorial units and additional information from the micro-region are used. The techniques used in this estimation are illustrated by an example in the field of waste management. In particular, it is an estimate of the amount of waste in individual municipalities. It is based on recorded waste production at district level and total waste management costs, which is available at a municipal level. To estimate the waste production, combinations of linear regression models with random forest models were used, followed by correction by quadratic and nonlinear optimization models. Such task could be seen as a multivariate version of inverse prediction (or calibration) problem, which is not solvable analytically. To test this approach, data for 2010 - 2015 measured in the Czech Republic were used.*

Keywords: *Data reconciliation, Random forest, Regression, Waste management, Optimization, Multivariate calibration, Inverse prediction*

1 Introduction

In real life optimization problems, there are some serious issues which could become bottlenecks even for the most sophisticated frameworks. One of the main sources of such issues is lack of reliable historical data (see [1]). The situation, when you have such data only for higher territorial units than you need, is particularly frustrating. From an optimization point of view, such problem can be seen as allocation problem. Such data are necessary for improvements, especially in fields with ongoing changes, where longterm predictions are needed. Example of such field is waste management [2]. In this case, we need to allocate waste quantities from district to municipality level. For this purpose, waste management costs (namely collection, manipulation, and overall expenditures), which are available on municipality level[1], will be used. Our approach is based on the idea that costs should be (linear) combination of waste quantities (namely sorted waste-SW, residual waste-RW, and bulky waste-BW), which is supported by Fig. 1. This could be seen as a regression model, but there is one important difference because, in regression, a variability of the dependent variable is supposed. In our case, there is variability present in predictors (when regression terminology used). Such problem, called calibration (inverse prediction), is well known and described [3]. The problem is that in our case, for every municipality (corresponding to one equation) there are three values to be estimated. From Frobenius theorem, such system has an infinite number of solutions. Thus, there is no method to solve such problem analytically and some heuristic approach is needed.

Fig. 2 shows the framework for our problem. In the beginning, only costs are known (for municipality level). In the first step, regression models for costs and waste quantities are built (on district level). Then, estimates of waste quantities, produced by models from the previous step (for municipalities), can be put into cost model (see the middle of Fig. 2). But since both cost and waste quantity models are created on a higher territorial unit, estimates of cost model coefficients need to be adjusted as well as waste quantities estimates. This process is called data reconciliation (see [4], where such an approach was used for waste-related data in the similar context), which comes from and is widely used in (petro)chemical industry [5].

[1]This data could be aggregated for district level

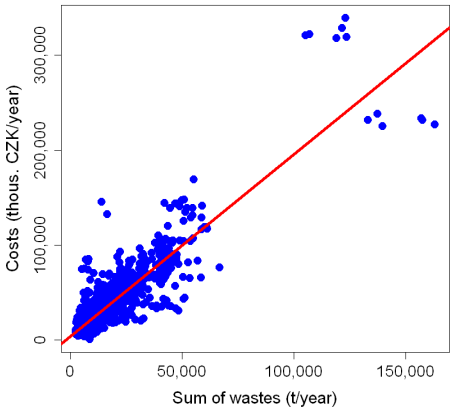

Figure 1: Dependency of Waste management costs (i.e. Total costs) on sum of waste quantities

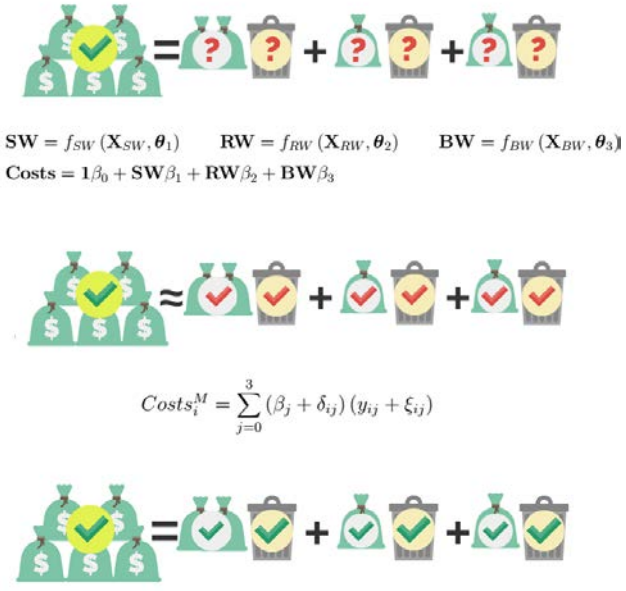

Figure 2: Ilustration of proposed framework (see Chapter 3 for notation)

2 Regression Models

2.1 Cost models

As mentioned in the previous section, proposed approach ensembles two groups of regression models. Cost models use collection and manipulation costs as dependent variable as well as their sum called Total. These were built on data from 2010 to 2014 with 2015 left as a test set. Linear model (LM) was selected for this purpose since linear dependency is assumed[2]. The results are shown in Fig. 3, LM Total performance is best as was expected since the Total variable is an aggregation of preceding two. From Fig. 3 it can also be seen that all of the predictors are significant, which is the basic assumption of the model (otherwise given waste would be collected and manipulated for free). Unfortunatelly, a problem with the normality of the residuals occurred, so Generalized linear model [6] (particularly Gamma regression) was used to deal with it. This approach helped to improve residuals a little, but classified some of the predictors as insignificant, which is seen as a more serious problem in this particular case. Last, models with transformations of dependent variables were examined, but the results were too bad to even consider their usage. Thus, for the rest of the work, LM Total model was used.

[2]Changing proportion of SW and RW should be considered, but enrichment by data quantity was taken as more beneficial

```
Residuals:
   Min     1Q Median     3Q    Max
-68556  -4671   -973   3890 102262

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 609.9663   683.1226   0.893  0.37212
SW            0.3324     0.1105   3.008  0.00269 **
RW            2.9452     0.1170  25.171  < 2e-16 ***
BW            2.5339     0.3328   7.614 6.05e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 ' ' 1

Residual standard error: 14360 on 1021 degrees of freedom
Multiple R-squared:  0.8361,    Adjusted R-squared:  0.835
F-statistic:  1736 on 3 and 1021 DF,  p-value: < 2.2e-16
```

	RMSE	MAE	MAPE	R^2
LM Collection	12,927	7,854	33.12	0.80
LM Manipulation	7,762	4,084	11,643.18	0.47
LM Total	14,334	8,521	30.88	0.84

Figure 3: Model summary for LM Total (left) and selected performance indicators for LM (right)

2.2 Waste quantity models

Modelling of waste quantities is more challenging task than previous one, largely because of the two reasons. First, there is a broader spectrum of possible predictors. Second, usually, there is no clear idea about the shape of the dependency between given waste and predictor. Forecasting in the waste management area was performed in [7], where spatially distributed hazardous production data were analysed. In this study, 41 socio-economical predictors were considered. Some of them showed a high (Spearman) correlation with the amount of waste. But there are also high correlations between some of the predictors (even after the removal of population and district area effects), which indicate possible multicollinearity in models based on these data. Because of this, principal component analysis (PCA) was done (see Fig. 4). When there was no clear idea of the shape of the dependency at the beginning, there is no idea about it after PCA. It worth a try to do LM on PCA so we can reduce the number of dependencies to be examined. Indeed, in some cases, it will be quite difficult to imagine linear dependency (maybe even arbitrary one).

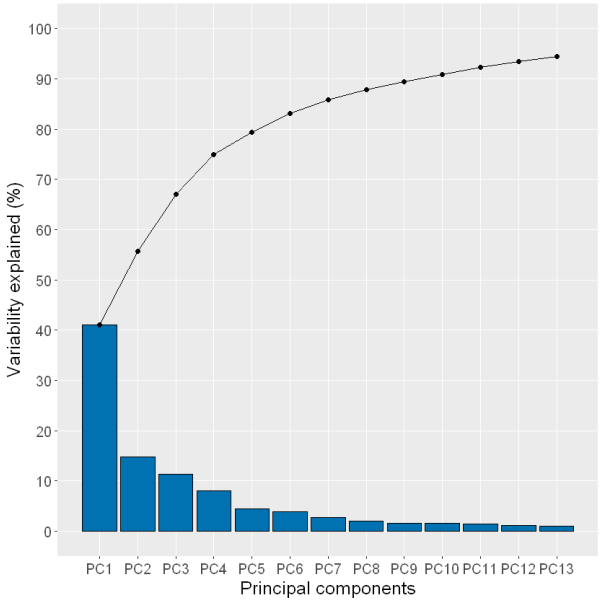

Figure 4: Pareto chart of variability explained by principal components

Some nonlinear shape could be fitted, but when the shape selection is not based on some fundamental idea, then its quality is uncertain. Thus, one of the machine learning models, Random Forrest (RF), was selected to compare its performance against PCA-LM models. It should be noted, that RF are generally considered a good choice for nonlinear models because of their good performance and little to no tuning needed [9]. Another important feature of RF is that it can estimate the missing values of predictors, which may be dangerous in some cases, but can be helpful too. Four datasets were created: M1 without predictors with more than 50% values missing (2 variables), M2.a from M1 excluding variables with one year of missing data (15 variables),

M2.b from M1 by excluding observations corresponding to variables removed in M2.a. Finally, M3 is composed just of variables which are present for all years for both district and municipality level. In M1, M2.a and M2.b missing data imputation was used for remaining variables.

Since RF is so called black box model, it will be examined more carefully. For these purposes, 7 training datasets were created by leaving out one year for testing (there is no problem with missing data for dependent variables from 2009 as in Cost models case). Performance of the best models is summarized in Table 1, while Fig. 5 supplies a graphical representation of these results.

Table 1: Selected performance indicators for RF models

	RMSE	MAE	MAPE	Rsq	Dataset
rf RW	870.97	570.23	7.76	0.98	M1
rf BW	811.74	382.58	38.23	0.85	M2a
rf SW	1,623.99	1,064.82	24.01	0.84	M1

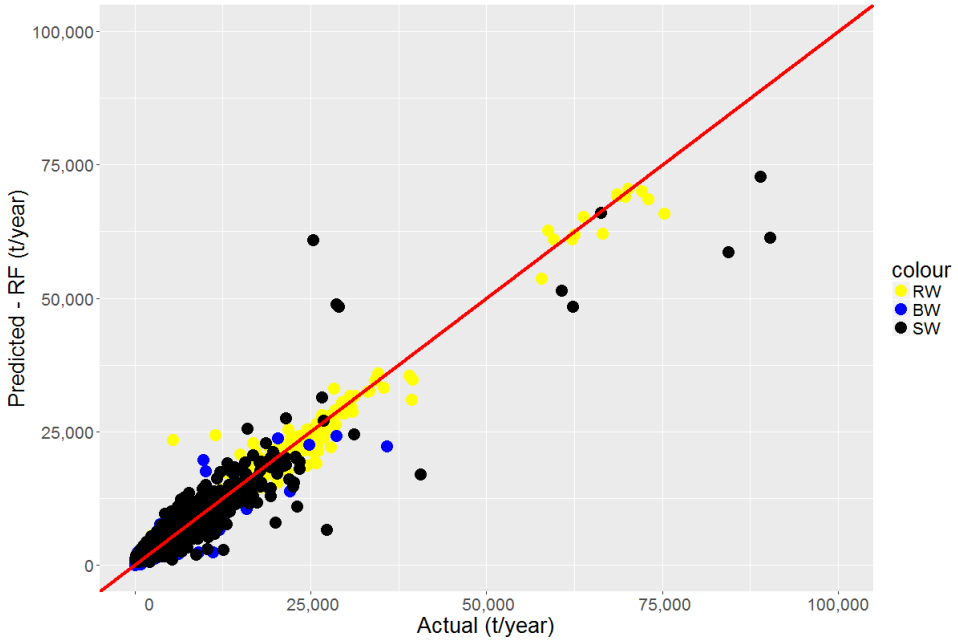

Figure 5: Comparision of actual and predicted values of SW, RW and BW

3 Data Reconciliation Models

In this section, three data reconciliation models will be introduced to adjust existing (regression) models. These adjustments should make the models more realistic. This should be achieved by satisfying conservation equations, namely law of mass conservation and (analogue of) law of energy conservation. For explanation purpose, equations (1), (2) and (3) will be used. From equation (1) it can be seen that the adjustments of both, model coefficients (β_j) and waste quantity estimates($\widehat{y_{i,j}}^M$) will be allowed. Equation (2) describes the fact that if the waste quantities from all of the municipalities of given district are aggregated, this value must be the same as waste quantity for the whole district. Equation (3) describes the same principle for costs. Since cost data were aggregated from municipality level to get data for districts, this equation will not be taken into account.

Sets

$i \in I$	index determining municipality
$j \in J$	index determining waste
$s \in S$	index determining district
$i \in O_s$	index determining municipality in s-th district

Parameters

$w^{\beta}_{j,j}$	weights of LM Costs model coefficients
$w^{Y}_{i,j}$	weights of $\widehat{y_{i,j}}^{M}$ estimates
$\widehat{y_{i,j}}^{M}$	estimate of j-th waste quantity of i-th municipality (per capita)
$y^{D}_{s,j}$	j-th waste quantity of s-th district
$\widehat{y_{s,j}}^{D}$	estimate of j-th waste quantity of s-th district (per capita)
$\widetilde{y}^{D}_{s,j}$	median of $y^{D}_{s,j}$ over S set
ν^{M}_{s}	costs for i-th municipality
ν^{D}_{s}	costs for s-th district
β_j	cost of j-th waste from LM Costs model
β_0	fixed costs from LM Costs model
κ_i	adjusted fixed costs
V_β	weight of $\delta_{i,j}$ for model 3.3
V_Y	weight of $\xi_{i,j}$ for model 3.3
POP_i	population in i-th municipality
$R^2_{P,j}$	variability explained by model for j-th waste ("pseudo" R^2)
$\lambda_{i,j}$	relative residual of waste quantity model

Variables

$\delta_{i,j}$	adjustment of j-th coefficient of LM Costs model in i-th municipality
$\xi_{i,j}$	adjustment of j-th waste quantity of i-th municipality

$$\nu^{M}_{i} = \sum_{j \in J} (\beta_j + \delta_{i,j}) \left(\widehat{y_{i,j}}^{M} + \xi_{i,j} \right) \qquad \forall i \in I, \tag{1}$$

$$y^{D}_{s,j} = \sum_{i \in O_s} (\widehat{y_{i,j}}^{M} + \xi_{ij}) \qquad \forall s \in S, \forall j \in J, \tag{2}$$

$$\nu^{D}_{s} = \sum_{i \in O_s} \nu^{M}_{i} \qquad \forall s \in S. \tag{3}$$

It is not advisable the weight adjustments of β_j and $\widehat{y_{i,j}}^{M}$ equally. Thus, these adjustments will be done separately. By doing this, we will find out the influence of these adjustments and set up appropriate weights for the final model.

While running models based on described ideas, some problems occurred. As the most important one is seen the appropriateness of waste quantity models. Models from 2.2 tend to overestimate waste quantities at the municipality level. Moreover, these models cannot capture smoother subtleties in predictors (in comparison with original scales). Because of this, such models are not suitable for data reconciliation. It should be noted, that some worsening of models by changing territorial level was expected. Such problem is well known from using standard regression models for predictions (model is valid only on the same domain as it was built on). Because the scope of this worsening was much bigger than expected, models for waste quantity per capita were created by the same approach as described in 2.2. RF models are still preferred since they are giving reasonable results (R^2_P from 0.4 to 0.9) compared to PCA-LM models (R^2 from 0.07 to 0.15). This adjustment will change the model just slightly because $\widehat{y_{i,j}}^{M}$ and ξ_{ij} just need to be multiplied by POP_i, then the original scale will be held.

3.1 Fixed quantities model

This model is the simplest one because since $\xi_{i,j}$ term is missing, (2) could not be held. Weights $w^{\beta}_{j,j}$ were introduced to allow the model to set up scales of adjustments for particular waste/coefficient. Median values (over the districts set S) were used for this purpose. Penalisation of adjustments of the quantities with good general prediction performance (R^2_P) was taken into account in $w^{Y}_{i,j}$. Moreover, $\lambda_{i,j}$ brings in the idea that if the waste quantity model works well for given district, there is no reason for big changes of estimates for municipalities in given district (both in terms of quantities and coefficients). Values of κ_i serve as the estimate of fixed costs.

$$V_\beta = \min_{\delta_{i,j}} \sum_{i \in I, j \in J} \delta_{i,j} w^{\beta}_{j,j} w^{Y}_{i,j} \delta_{i,j} \tag{4}$$

s.t.

$$\nu_i^M = \kappa_i + \sum_{j \in J} \widehat{y_{i,j}}^M POP_i(\beta_j + \delta_{i,j}) \qquad \forall i \in I, \tag{5}$$

$$\kappa_i = POP_i \frac{\mathrm{card}(S) \cdot \beta_0}{\sum_{i \in I} POP_i} \qquad \forall i \in I, \tag{6}$$

$$w_{j,j}^{\beta} = \widetilde{y}_{s,j}^D \qquad \forall j \in J, \tag{7}$$

$$w_{i,j}^Y = \frac{R_{P,j}^2}{\lambda_{i,j}} \qquad \forall i \in O_s, \forall j \in J, \tag{8}$$

$$\lambda_{i,j} = \frac{100|\widehat{y_{s,j}}^D - y_{s,j}^D|}{y_{s,j}^D} \qquad \forall i \in O_s, \forall j \in J. \tag{9}$$

3.2 Fixed coefficients model

Fixed coefficients model is more complex since constraint (2) is used (coefficients correspond to unit costs). There is one more constraint present in the model. Constraint (13) ensures non-negativity of waste quantities, which is completely natural.

$$V_Y = \min_{\xi_{i,j}} \sum_{i \in I, j \in J} \xi_{i,j} w_{j,j}^{\beta} w_{i,j}^Y \xi_{i,j} \tag{10}$$

s.t.

$$\nu_i^M = \kappa_i + \sum_{j \in J} (\widehat{y_{i,j}}^M + \xi_{i,j}) POP_i \beta_j \qquad \forall i \in I, \tag{11}$$

$$y_{s,j}^D = \sum_{i \in O_s} (\widehat{y_{i,j}}^M + \xi_{i,j}) POP_i \qquad \forall s \in S, \forall j \in J, \tag{12}$$

$$(\widehat{y_{i,j}}^M + \xi_{i,j}) POP_i \geq 0 \qquad \forall i \in I, \forall j \in J, \tag{13}$$

all together with the equations (6),(7),(8),(9) from fixed quantities model. Both mentioned models are the tasks of the quadratic programming. The constraints of the model are linear, and the objective function is strictly convex (the quadratic objective coefficient matrix is positive definite). This means that the tasks have only one unique solution that is the global optimum.

3.3 Variable quantities and coeffiecients model

Although this model is just slightly different when compared to the previous models (it is basically the combination of two), some inconvenience occurs. In constraint (15) multiplication of variables ($\xi_{i,j}$ and $\delta_{i,j}$) pops out, which means that the model is no longer quadratic programming problem (as it was in the previous two cases), but the general nonlinear programming problem. This fact implies unpleasant consequences as no guarantee to reach the global optimum.

$$\min_{\xi_{i,j} \delta_{i,j}} \frac{1}{V_Y} \sum_{i \in I, j \in J} \xi_{i,j} w_{j,j}^{\beta} w_{i,j}^Y \xi_{i,j} + \frac{1}{V_\beta} \sum_{i \in I, j \in J} \delta_{i,j} w_{j,j}^{\beta} w_{i,j}^Y \delta_{i,j} \tag{14}$$

s.t.

$$\nu_i^M = \kappa_i + \sum_{j \in J} (\widehat{y_{i,j}}^M + \xi_{i,j}) POP_i (\beta_j + \delta_{i,j}) \qquad \forall i \in I, \tag{15}$$

$$y_{s,j}^D = \sum_{i \in O_s} (\widehat{y_{i,j}}^M + \xi_{i,j}) POP_i \qquad \forall s \in S, \forall j \in J, \tag{16}$$

$$(\widehat{y_{i,j}}^M + \xi_{i,j}) POP_i \geq 0 \qquad \forall i \in I, \forall j \in J, \tag{17}$$

all together with the equations (6),(7),(8),(9) from fixed quantities model.

4 Results

All of the models from Section 2 were build using R programming environment [8]. Outputs of these models (LM Total model coefficients and RF models predictions) were used as inputs for data reconciliation. Results from model 3.1 and 3.2 (i.e. $\delta_{i,j}$ and $\xi_{i,j}$) were then used as initial values for model 3.3, which provides final estimates. Such approach should help to find a better solution in case of nonlinear programming problem. Optimization part was done by using AMPL software with CPLEX (for models 3.1 and 3.2) and MINOS solvers. The results were compared with actual data from 2014.

The biggest issue of the proposed framework is that waste quantity models cannot adapt to municipality level, as already mentioned. Improvement of these models is essential to get meaningful results. Some data reconciliation on waste quantity models could lead to such improvement, but for this, "white box" model with sufficient performance is needed.

Moreover, Fig. 6 shows that although most of the residuals (before data reconciliation) are spread around 0, there is some indication, that residuals are negatively skewed (extreme values were excluded for vizualisation purpose). It means that L^1 norm could be considered in the objective function. The same is true for waste quantity adjustments - Fig. 7 (just RW adjustments are not strongly skewed). However, this leads to even more extreme changes, which is not desirable. Fig. 8 shows that skewness of unit cost adjustments is smaller when compared to waste quantity adjustments.

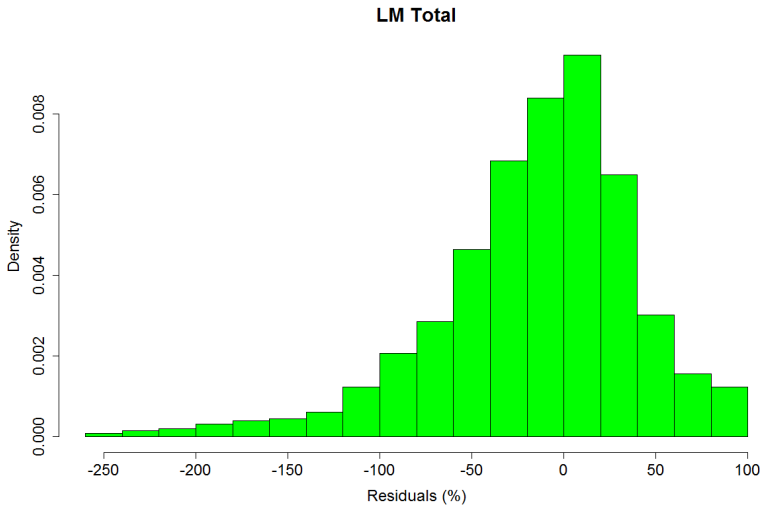

Figure 6: Histogram of LM Total residuals (%) on municipality level before data reconciliation (more than 200 values $\leq$ -250 excluded)

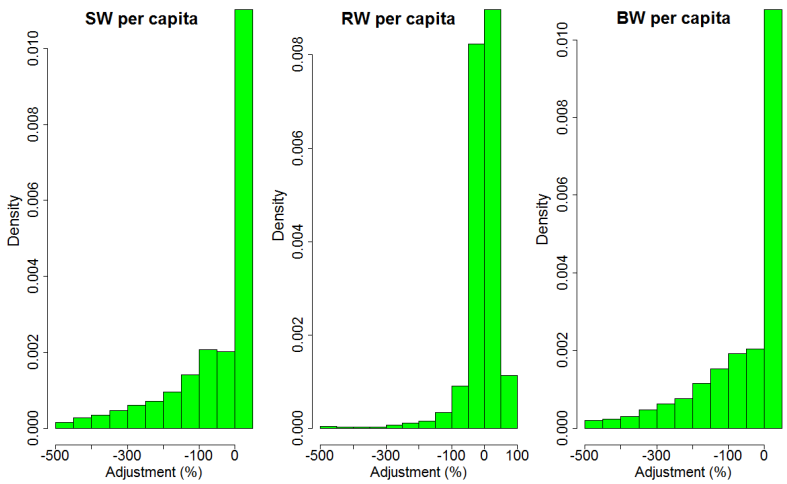

Figure 7: Histograms of per capita waste quantity adjustments (%) after data reconciliation (values $\leq$ -500 excluded, namely 584 for SW, 63 for RW and 491 for BW)

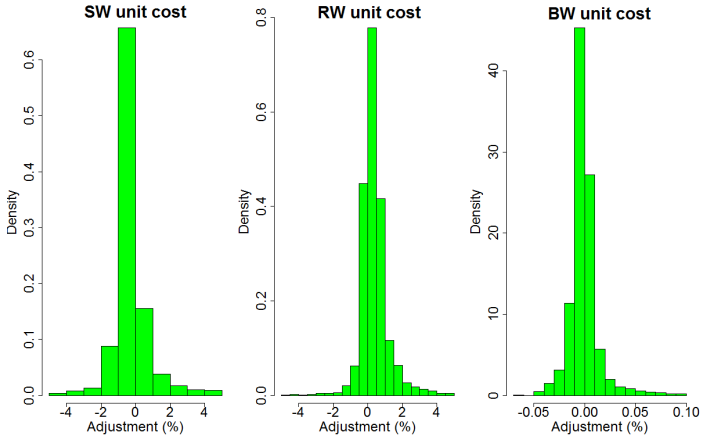

Figure 8: Histograms of unit cost adjustments (%) after data reconciliation (265, 156 and 110 extreme values excluded)

5 Conclusion

This article describes the heuristic approach for solving multivariate inverse prediction problem in waste management. It is basically based on usage of linear regression models for the Total cost of waste management. Random Forest models were used to estimate the amount of waste since traditional methods were not able to provide sufficient results. Further, the article devoted to the balancing of the estimated values from the mentioned models. Two models of quadratic programming were created, especially for the amount of waste and especially for the Total cost. The results from these models were used in the final nonlinear model, which includes both the Total cost and the amount of waste. As a result, municipal waste data was generated. The results which were evaluated on historical data for 2014 are summarised in Fig. 6-8. For future work, it would be necessary to develop waste quantity models, which are less sensitive to change of the scale. This seems like a very difficult problem and it is not clear whether it would be possible to reach significant improvement in this way.

Acknowledgement: The authors gratefully acknowledge the financial support provided by the project Sustainable Process Integration Laboratory SPIL, funded as project No. CZ.02.1.01/0.0/0.0/15 003/0000456, by the Czech Republic Operational Programme Research and Development. The related research was also supported by solution grand FV18-50 science Fund of the FME 2018 at Brno University of Technology.

References

[1] Smejkalová, V., Šomplák R., Nevrlý V., Pavlas M.: Design and decomposition of waste prognostic model with hierarchical structures. In: R. Matousek (ed.), In MENDEL, (MENDEL 2018), vol. 24, No. 1, Brno University of Technology, Brno (June 18), ISSN 1803-3814. Manuscript submitted for publication.

[2] Beigl, P., Lebesorger S., Salhofer S.: Modelling municipal solid waste generation: A review. Waste Management **28**(1) 200–214 (2008). DOI https://doi.org/10.1016/j.wasman.2006.12.011

[3] Osborne, C.: Statistical Calibration: A Review. International Statistical Review / Revue Internationale De Statistique **59**(3), 309–36 (1991). DOI 10.2307/1403690

[4] Nevrlý, V., Šomplák R., Popela P., Pavlas M., Osička O., Kůdela J.: Heuristic challenges for spatially distributed waste production identification problems. In: R. Matousek (ed.), In MENDEL, (MENDEL 2016), vol. 22, No. 1, pp. 109–116, Brno University of Technology, Brno (June 16), ISSN 1803-3814

[5] Narasimhan, S., Jordache, C.: An Intelligent Use of Process Data, *Data Reconciliation and Gross Error Detection*. Gulf Publishing Co., Houston (2000)

[6] McCullagh, P., Nelder, J. A.: *Generalized linear models*, second edn. Chapman & Hall, New York (1989)

[7] Pavlas, M., Šomplák R., Smejkalová V., Nevrlý V., Szásziová L., Kůdela J., Popela P.: Spatially distributed production data for supply chain models - Forecasting with hazardous waste. Journal of Cleaner Production **16**, 1317–1328 (2017). DOI 10.1016/j.jclepro.2017.06.107

[8] R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

[9] Kalmar, M., Nilsson, J.: The art of forecasting – an analysis of predictive precision of machine learning models. PhD thesis, Uppsala University (2016)