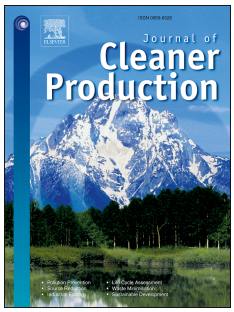
Accepted Manuscript



Forecasting multiple waste collecting sites for the agro-food industry

Julio Montecinos, Mustapha Ouhimmou, Satyaveer Chauhan, Marc Paquet

PII: S0959-6526(18)30788-1

DOI: 10.1016/j.jclepro.2018.03.127

Reference: JCLP 12385

To appear in: *Journal of Cleaner Production*

Received Date: 1 November 2017

Revised Date: 11 March 2018

Accepted Date: 12 March 2018

Please cite this article as: Montecinos J, Ouhimmou M, Chauhan S, Paquet M, Forecasting multiple waste collecting sites for the agro-food industry, *Journal of Cleaner Production* (2018), doi: 10.1016/j.jclepro.2018.03.127.

Forecasting Multiple Waste Collecting Sites for the Agro-food Industry

Julio Montecinos[a,*], Mustapha Ouhimmou[a], Satyaveer Chauhan[b], Marc Paquet[a]

[a]*Automated Manufacturing Department, cole de technologie suprieure (ETS), Montral, Qubec, Canada*
[b]*Supply Chain & Business Technology Management, Concordia University, Montral, Qubec, Canada*

Keywords: Waste management, Agro-food Industry, Forecasting, Time series, Theil-Sen.

*Corresponding author
Email addresses: julio.montecinos@etsmtl.ca (Julio Montecinos), Mustapha.Ouhimmou@etsmtl.ca (Mustapha Ouhimmou), satyaveer.chauhan@concordia.ca (Satyaveer Chauhan), Marc.Paquet@etsmtl.ca (Marc Paquet)

Forecasting Multiple Waste Collecting Sites for the Agro-food Industry

Julio Montecinos[a,*], Mustapha Ouhimmou[a], Satyaveer Chauhan[b], Marc Paquet[a]

[a] *Automated Manufacturing Department, École de technologie supérieure (ETS), Montréal, Québec, Canada*
[b] *Supply Chain & Business Technology Management, Concordia University, Montréal, Québec, Canada*

Abstract

The agro-food industry wastes tons of oil and grease not suitable for immediate consumption. Their collection mostly relies on the experience of managers and this results in inaccurate visits by truck drivers and operations teams. Indeed, the measurement of by-products waste is complex and thus information is imprecise, making the collecting operations inefficient. In this paper, we propose a model that forecasts the daily input of thousands of industrial and commercial sites of the agro-food industry based on historical data. The algorithm rejects errors and mistakes in the routing-collection-measuring process. In our model, the site container capacity is known and remains constant. The main contribution of this study is to propose a model based on the Theil-Sen constrained regression (Theil-Sen CR) that rejects errors and outliers to simplify the forecast of future collections. We apply this method to a real case study and compare its performance at different collecting sites. The forecasting error is significant compared to Linear Regression (LR). We have calculated, for our industrial partner, based on 12.2 km between sites and a fleet of 200 trucks, a potential reduction of 940 tCO_2 equivalent per year.

Keywords: Waste management, Agro-food Industry, Forecasting, Time series, Theil-Sen.

1. Introduction

Valuable waste is the waste that can be transformed into a commercial resource. Within this category of waste, food oils are attracting the attention of the industry, as they can be used as an infant dietary supplements (Bialy et al., 2011) or as construction material (Nadeem et al., 2017). They have been also used as biodiesel from frying oil (Lam et al., 2017) or trapped grease (Tu and McDonnell, 2016). In addition, used cooked oil has been seen to be effectively collected in a sustainable way (Vinyes et al., 2013). In this study, we focus on valuable wastes generated by restaurants, grocery stores, supermarkets and slaughterhouses, which typically use cooked oil and animal fats (Tran et al., 2016). Fat, oil and grease are commonly abbreviated as FOG. An example of the supply chain behind valuable waste recycling can be seen in Figure 1. In this figure, commercial and industrial waste is stored in special containers where specialized trucks collect these by-products to be transported to a processing plant. The output of this plant consists of several raw materials to be re-injected in the common supply chain for other industries or farms closing the recycling supply chain loop. Valuable waste recycling must deal with the existence of inadequate fixed or mobile containers (Andrews et al., 2013), gross schedules (Zsigraiova et al., 2013), bad quality of roads for general waste collection, vehicle routing and planning, as well as inadequate or insufficient vehicle fleet (Fitzgerald et al., 2012). Specialized tools can help to plan the routing and the tracking of specialized equipment, but they are not adapted to the oil recycling collection problem because of the lack of precise updated information from the sites, which is different from the general waste collection as in Han et al. (2015). Indeed, the initial routing plan does not follow changes in waste generation

*Corresponding author
Email addresses: julio.montecinos@etsmtl.ca (Julio Montecinos), Mustapha.Ouhimmou@etsmtl.ca (Mustapha Ouhimmou), satyaveer.chauhan@concordia.ca (Satyaveer Chauhan), Marc.Paquet@etsmtl.ca (Marc Paquet)

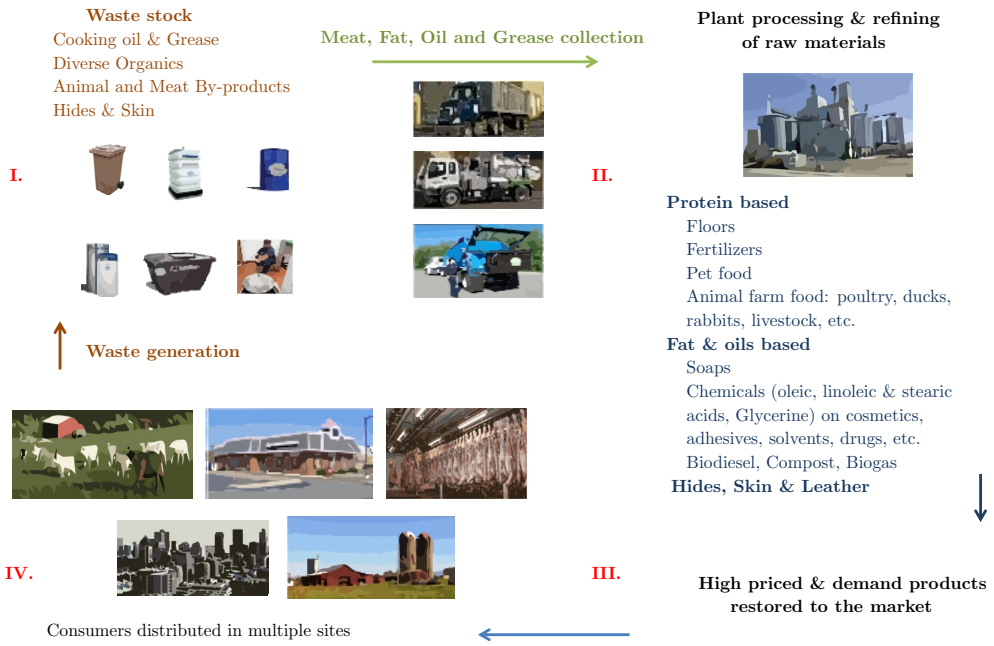


Figure 1: Reverse Logistic of Recycled valuable waste recycling

Diagram of a recycling and treatment supply chain. Commercial and industrial waste is stored (1) specialized trucks transport the waste to a processing plant (2), the plant refines on demand raw materials to be sold (3) to other industries or farms (4). Source: Sanimax Inc. and Literature Review

because it is usually modified to respond to last-minute requests, to deal with bad weather conditions, traffic conditions, container problems, contaminated production and/or pumping difficulties. This problem, added to the difficulties in the measurement of the actual waste deposited in the tanks, impedes the use of common even time series models to better forecast the daily waste of production sites. Even the effective volume of containers is difficult to measure as different sites have permanent or temporary containers adapted to their operations. There are several publications examining all the technical aspects of recycling, but only a few of them have been focused on studying the best mechanisms to address the recycling collection problem, such as (Guabiroba et al., 2017).

1.1. Waste Management Operations

Waste management consists of "the whole set of activities related to handling, treating, disposing or recycling the waste materials" (Ramos et al., 2013). Recycling is usually intended when it is profitable (Tonjes and Mallikarjun, 2013) and when the quantifiable damage to the environment from recycling (Arena and Di Gregorio, 2014) is lower than the extraction from natural sources. Life cycle assessment (LCA) combined with material and substance flow analysis can be used in the quantification of such environmental impacts (Laurent et al., 2014). Reverse logistics attempt to ensure that materials are returned to be reused or reconditioned (recycled). It consequently requires management actions and plans for reprocessing agro-wastes or perished foods from consumers. For example, Sharma et al. (2017) explores the diverse performance indicators related to a green supply chain management. Current research like Singh et al. (2018), explore the importance of supply chain collaboration and information flow in reverse logistics achievement.

Most of the collecting operation consists of gathering materials subject to the capacity of adapted vehicles and on-site facilities. Collecting operations are linked

to the vehicle routing problem (VRP) in the sense that collecting companies have to visit several independent clients in a finite rolling horizon (Aksen et al., 2012). Moreover, clients impose time windows, and this results in unused vehicle capacities. Prior studies in waste management also show that specialized vehicles and multi-waste collection can have a huge impact on management costs (Teixeira et al., 2004). Long collection delays are unacceptable and companies are penalized because waste becomes a safety hazard, which is translated into penalties for the recycling companies. In order to be profitable, companies engage in agreements with a huge number of collection sites but this practice is not always beneficial as the variability among clients' productions increases too. Companies in charge of transportation try to minimize the operations costs collecting separated types of waste, when this is possible. Examples can be seen in Teixeira et al. (2004). We argue that the forecasting of FOG waste based on noisy historical data from multiple sites of the agro-food industry has not been sufficiently addressed in the literature. To our knowledge filtering outliers with the intention of correcting mistakes introduced in the collection process and also smoothing the same data using robust regressions has not been completely treated in the context of estimating good periods for collections.

1.2. Problem Statement

An automatic robust forecasting of multiple sites is important for choosing profitable collection sites and truck routing. Until recently, the industry did not see the need for using special containers; recycling companies were working under limited protection protocols and without proper measurement and instrumentation to account for precise records (Faria et al., 2016). Tagging and the geolocalization of bins and sites are useful for fast response routing and traceability, but they are insufficient for forecasting, as quantities can vary. Telemetry and new interconnecting technologies facilitate the tasks of acquiring real-time data (Hannan

et al., 2015), from different sites, but they increase costs and dependability as every site must keep measuring-communication systems. There are at least two ways to reduce these costs: increase the connectivity and instrumentation on the collecting trucks (Arebey et al., 2010). Another alternative could be to use remote sensors and public cellular networks or the networks of a smart city that can usually be shared with different users (Folianto et al., 2015). Smart cities will be provided with image processing, high-speed networks and environmental measurement of gases (Hancke et al., 2013), which could also be used to monitor the bins and recycle waste (Gutierrez et al., 2015).

As remote sensing and partnering for sharing real time information are not common practices, the industry of valuable waste collection looks for historical information to plan these operations. In practice, most of these waste collecting operations are supported by sampling and forecasting to determine static and dynamic collections (Mes, 2012; Mes et al., 2014) policies. The sampling is done by three methods: periodic sampling, on-call sampling (random) and neighboring association sampling (if a near point is sampled, then close points are also sampled). For its simplicity and reduced costs, periodic sampling is the common paradigm. It consists of a periodic scheduled routing (Campbell and Wilson, 2014) complemented with surveillance and pulled calls to the collection sites or industrial partners. "Over-sampling" (excessive monitoring) is also penalized as it distracts small businesses from their duties and increases their costs. Based on this information, the corrective actions affect the VRP established and force last-minute changes. Another important factor in industrial waste collection forecast is that clients appear or disappear dynamically, creating incertitude in local areas. Then time series analysis needs to cope with uneven-spaced series, missing information or inconsistent noise due to human errors. It is interesting to note that special events can induce spikes in the records.

It has been noted that waste sites can be modeled using ARIMA (Box-Jenkins), GARCH or Exponential smoothing techniques (Huddleston et al., 2015). This is appropriated when clients remain in stable operations for a long time and sampling can be done periodically, generally with high frequency. In consequence, sampled time series can be approximated as a stationary series and thus trends & seasonality could be easy to model, if any. Then forecasting time series techniques suggest the use of Seasonal autoregressive integrated moving average (SARIMA) models, artificial neural networks (ANN) and the reconstruction of the space state model (SSM) for this specific type of forecast. In fact, most vendors and developers of traditional statistical and mathematical suits (SAS®, STATA®, SPSS®, STATGRAPHICS®, R (Team, 2004)) have developed libraries for dealing with data mining, using artificial intelligence, deep learning, statistical inference, etc. The automatic forecasting of time series is a debatable subject of the past decade, as the idea is to simplify common planning tasks (Yan, 2012). The treatment of numerous data time series that are constantly changing is not easy with complex methods. One of the best known packages that attempts this is "Forecasting" for the R Language (Team, 2004). This package treats state space models, exponential smoothing methods and forecasting with ARIMA models, considering stationary and non-stationary data (Hyndman and Khandakar, 2008). However, non-stationary treatments are not automatic functionalities of the software and thus users must parameterize it to conduct initial transformations to shape the residual error to be normal. The forecast must be transformed back in the departure space. Another interesting approach for forecasting related time series is called Ensemble Time Series classification (Bagnall et al., 2012). The project Prophet (Taylor et al., 2017) works by weighting or combining different forecasting models to improve the individual accuracy and it should work for several data of similar known characteristics. A common approach could be to correct the data and try seasonal and special events or holiday adjustments to test several

models and parameters seeking to satisfy the best of fit measure, penalizing the over-parameterization. Any forecast would need to reapply the correction for special events and holidays on specific days or on an aggregated level, such as day-week-month, etc. Measuring the appropriateness and quality of the model is another challenge.

In summary, the case of valuable waste collection for industrial and commercial clients does not have simple characteristics for profitability analysis. Usually, historical sampled data is biased by mistakes and unknown events. Cannibalistic behavior among clients can happen and be hidden as the forecast remains done for individual/independent clients and without geographical awareness. The contribution of this paper is to propose a method suitable for forecasting of daily valuable waste based on the historical data of the agro-food industry. The algorithm filters errors in the collection-measuring process. The results, based on the average, have a maximum potential of 32.6 % reduction in the number of visits, which could imply a decrease in transportation costs and gas emissions. We have calculated, based on 12.2 km between sites and a fleet of 200 trucks, a potential reduction of 940 tCO_2 equivalent per year.

The remaining of the article is organized as follows: Section 2 describes the method used. Section 3 presents and analyses the numerical results and Section 4 presents the conclusions and future research.

2. Materials and Methods

In this paper, we study real data on weight and collection dates from several collections operations in North America provided by our industrial partner. Further information regarding the collection sites is proposed in subsection 2.2 Case Study. We show that LR is inadequate to forecast the filling ratio of FOG waste at different sites. With this information, we propose a better estimator that

circumvents its limitations keeping good forecasting. Then, we show that current periodic visits are inefficient for collecting waste in comparison with the new estimator and known capacities.

2.1. Robust Regression

The use of linear least squares is usually considered for estimating the parameters (m, b) of $y = mx + b + \epsilon$, with ϵ the deviations original data. The results of this regression rely on assumptions such as linearity, homoscedasticity, and the absence of measurement errors (or the independence of these errors), among others. In different applications, outliers have excessive influence on the results. In any case, at least some assumptions are needed: The outliers must be few and there should be enough points for defining an underlying structure with linear alike characteristic.

Robust regression methods such as the Theil-Sen regression or the Kendall line-fit method are devised to avoid limitations of the least squares method because these methods consider errors in the dependent variable. Research like in Wilcox (1998) indicates that both methods can be very efficient. In the definition of the Theil-Sen regression, m is the median of the slopes $\frac{(y_j - y_i)}{(x_j - x_i)}, \forall (x_i, y_i)$ with $1 \leq i < j \leq n$ and $x_i \neq x_j$; and b is the median of $(y_i - mx_i), 1, \ldots, n$. This estimator has good properties like invariance and biasness (Sen, 1968). Some implementations of this estimator rely on sampling to reduce the number of calculations when "n" is too big or in choosing more probable points to accomplish smoothing periodical data as in Hirsch et al. (1982). In particular in Scikit-learn (Pedregosa et al., 2011) for Python, which uses sampling for long series of data, the spatial median is calculated for all the couples of points in the sample. The spatial median is estimated using interactive least squares. Another useful estimator is the "Pairwise mean scale estimator" (PMEE), a robust non-parametric estimator (Tarr et al., 2012) that can be adjusted to estimate the

standard deviation of Gaussians. In the proposed algorithm, it is used to bound the outliers, as can be seen in Appendix B in supplementary material.

For the waste collection process, time and collections (X and Y) are proportional. As storage continually increases in the industrial collection of products, it is possible to argue that any straight line that attempts to follow that increment cannot have any negative slope "m" and that the intercept should remain near the origin when the collection is completed, but can't be negative and usually not zero. As the problem of determining the best date for collecting depends on the increase of the product from the intercept, an estimation that results in a negative slope "m", despite its goodness of fit, is discarded by the industry in favor of the model $y = mx$. But this regression can still be affected by outliers and the intercept is neglected. An example can be seen in (supplementary material Appendix A).

As previously mentioned, the median of slopes in Theil-Sen regression is an unbiased estimator resistant to some errors and outliers. However, a large number of outliers can increase the number of slopes with negative values, or intercepts that exceed the container's capacity. Let's consider that the time between collections is "x" and the collections weight is "y". In this situation, it is important to emphasize the couple of points that match an LR with positive slopes and intercepts in the range of the most probable remaining material after collection, i.e. the term "b" in the regression. To avoid this disadvantage, we propose to begin with the pairs of points (x_i, y_i) and (x_j, y_j) (with $1 \leq i < j \leq n$ and $x_i \neq x_j$) that follow this principle and add progressively other points' pairs (no outliers) if the final regression parameters respect the positive slope and intercept in the range of the plausible remaining material. With this procedure the automatic forecasting could really help to choose the adequate timing for picking profitable sites and influence the trucking routing.

Our implementation finally considers only a positive slope and a small positive intercept (equivalent to no more than 15 % of the container size (or the max

collection)). The random selection of couples of points is not affected by any restriction. We note that similar slope values $m_i \approx m_j$ can be related to widen different intercept $(y_i - m_i x_i) \neq (y_j - m_j x_j), i \neq j$ and the opposite is also possible $(y_i - m_i x_i) \approx (y_j - m_j x_j), m_i \neq m_j, i \neq j$. The regression parameters $(m_{\mathrm{TS}}, b_{\mathrm{TS}})$ must be compatible with the process having the property $(0,0) < (m_{\mathrm{TS}}, b_{\mathrm{TS}}) < (\infty, C)$ with C a small constant, representing the max amount known to have been left in a container (approx. 15 % of the container).

We can see the possible restrictions as if we have in one extreme a Theil-Sen regression that forces the passage by the origin $b = 0$, i.e. always considering an $(x_i, y_i) = (0,0)$, followed by the original Theil-Sen regression that first estimates the slope and then the intercept, and finally our constrained version that finds both $(0,0) < (m_{\mathrm{TS}}, b_{\mathrm{TS}}) < (\infty, C)$, without restricting the points in consideration. We consider a sample that also includes few negative slopes and wide intercept values to reduce any bias and that gets errors close to form a normal distribution for the non-outliers' points. We note that ideally $(m_{\mathrm{TS}}, b_{\mathrm{TS}})$ should be as efficient as the least square LR (for the case without outliers).

2.2. Case Study

The data for this study was provided by Sanimax Inc. The company estimates the recuperation of 2 billion kilograms of by-products that will be renovated and converted into animal feed, pet food, soaps, leather, lubricants, cutting oils, paint, rubber, cosmetics, perfumes, inks, adhesives, solvents, antifreeze, fertilizers, biofuels, etc. Nevertheless, its operations force trucks to continuously pool customers across the country in very long tours. The filling of the containers and tanks is uncertain and the company visits its clients on a periodical basis to prevent emergency client overflow calls. As shown in (supplementary material Figure A.1), the time window for more profitable collections induces a preference at the time of solving the trucking problem.

The waste collection problem is difficult as the forecast of the filling ratio is needed to determine the longest time for collection, which is never reached in practice, and to construct a schedule for the collection periods that will improve the touring planning. We assume that most collection sites do not have increasing product collections over time as most commercial and industrial sites have a constant maximum container capacity already defined and their space is limited. With a carefully developed transportation plan, the truck traveling time, the fuel consumption and truck emptiness will be reduced.

From more than 20,000 collection sites in the east-south coast of Canada and the east-north coast of the United Stated of America and under anonymity of the clients, we choose those 116 making frequent visits and 5-year historical data. The data currently stored in the Enterprise Resource Planning System was collected from truck-collections operations and it was primarily used for billing.

2.3. Statistical Procedure

We use the following steps to obtain the numerical results.

1. Estimate the Linear Regression, Constrained Theil-Sen Regression, Median Rates, Exponential Smoothing and Moving Average for "FOG collection weight vs time" (the ratio) considering several independent sites.
   The data plots, the Box-Cox plots and the Histograms graphs are shown in the supplementary material of the paper. Using these graphs, the "Odd cases" of Linear Regression can be spotted.
2. We perform the following two tests and compare the outcomes:
   (a) Test 1: LR Comparison: The Linear Regression is compared to Theil-Sen Constrained, Median Rates, exponential smoothing, moving average, etc., on the goodness of fit.

(b) Test 2: Short Run Forecast Comparison: It considers few samples in advance for the forecasting accuracy. Particular parameters are optimized for the estimations.

(c) We compare these individual estimations using the "Root mean square error" (RMSE) and the "Mean absolute error" (MAE).

3. We revise the site estimations' differences considering 0.05 % of significance, using the following parametric and non-parametric tests:

   (a) The "Two sample t-test" on the normalized data. This test assumes normality.

   (b) The "Sign Test", that compare differences on paired observations. It does not assume a normal distribution

   (c) The Wilcoxon signed-rank test, to compare observations on the rank. It does not assume a particular data distribution.

4. Finally, estimate the maximum transportation reduction by using the suggested Theil-Sen Constrained regression compared to the current period used.

3. Numerical Results

The initial test considers the goodness of fit of past data using the proposed method, and compare it to the goodness of fit of common industrial methods, like LR. We note that LR can have a very high goodness of fit with high intercept levels or even have negative (or almost zero) filling ratios. Our new method should have comparable performance against the LR, but it is not mandatory as the LR can be odd. The comparison consists of 116 time series selected from the 20,000 collecting sites with more visits.

It is possible to see, in the supplementary material in Appendix E, several plots of time series from different collection sites. In every plot, it is possible to

compare the Theil-Sen CR with the LR. In some cases, the LR presents very counterintuitive results, big intercepts or negative slopes, because of outliers in the data. Some cases can be enumerated: We have 2 cases (cases "30" and "54") where the LR gives negative slopes and in many other cases even a parallel slope with a big intercept with respect to past collections, like in case "16", "76" and "78". In one case, the "28", the Theil-Sen CR and LR converges to an almost zero slope because of the existence of several "aligned" points, which indicates the limitations of the method with too noisy data. In most cases, the Theil-Sen CR passes through the intercept of both central tendencies lines, keeping the intercept very close to the origin.

Two different tests have been considered to compare forecasting. On the one hand, a comparison with the LR: Only the central tendency will be analyzed, which is an important case for the industry. In another test, we will run a short forecast that could provide better results because these methods attempt to follow the evolution of the ratio to compare the goodness of fit. Both tests can be seen in the supplementary material Appendix C, as Table C.2, Table C.3 and Table C.4. Those tables present the errors for LR, Theil-Sen CR, median rates (MR), moving average (MA) and Simple exponential smoothing (SES).

3.1. Test 1: LR Comparison

It is possible to compare the goodness of fit for both regressions based on the historical 4 last collections using the "Root mean square error" (RMSE) and "Mean absolute error" (MAE) indicated in Table C.2, Table C.3 and Table C.4. These tables also present the case when the LR gives odd results "Odd LR" (Table C.2, "Odd LR" both parts) and the case when LR gives plausible results (Table C.3 and Table C.4, "No Odd LR"). There are 3 sets of figures: Full data, Odd LR and No Odd LR to compare. The full data figures and the Odd LR figures have too much in common. The Appendix D, in supplementary material, presents

the histograms and Box-cox plots for the samples. The histogram of the errors can be seen in Figure D.2, Figure D.4, Figure D.6, Figure D.8, Figure D.10 and Figure D.12. These errors do not have an identifiable distribution, but they have long tails as can be seen in Table C.2, Table C.3 and Table C.4. The Normalized Box-Cox plot can be seen in Figure D.3, Figure D.5, Figure D.7, Figure D.9, Figure D.11 and Figure D.13. The Box-Cox normalized plot allows us to compare central tendency and variance more easily. We note nevertheless that the normalization process applied to the series fails to pass strong normalization tests because of the long queues in the data. In general, Theil-Sen CR is comparable to LR for the full data analysis and both behave slightly better than SES and MA. As it is possible to compare from these figures, "Odd LR" shows longer tails than "No Odd LR".

According to Figure D.3, Figure D.5, Figure D.7, Figure D.9, Figure D.11, Figure D.13 and Table 1, showing only the means and medians for both RMSE and MAE, the "Odd LR" gets worst results, which could be an indication that LR parameters are misled by outliers. Also, the (normalized) skewness improves for the case "No Odd LR". The developed Theil-Sen CR gets almost similar results as LR and slightly better than SES, MR and MA. This could be indicating that uncertainty is difficult to forecast and that a global regression, when valid, is enough. Theil-Sen CR seems to be slightly better but not significant. In Table C.2, Table C.3 and Table C.4 it is also shown which forecasting method achieves the lowest error in every case in the column "Min col". A comparison in percentages with respect to "LR MAE" is provided to show how big the differences are presented across sites and how widely they spread across the forecasting methods. Therefore, an undiscriminating use of LR as an estimator could induce considerable errors for the period estimation as well as for the truck routing problem.

Table 1: RMSE and MAE compared with the mean, median and skewness able (no transformation)

	RMSE					MAE				
	Linear Regression	Theil-Sen Constrained	Exponential Smoothing	Median Rates	Moving Average	Linear Regression	Theil-Sen Constrained	Exponential Smoothing	Median Rates	Moving Average
Odd										
Means	921.5	1,131.1	1,263.5	1,286.8	1,309.7	755.4	927.7	975.3	1,246.2	1,286.9
Medians	145.7	157.4	162.7	167.3	159.1	108.3	134.6	147.9	162.4	158.6
Skewness	4.5	4.1	4.4	4.1	3.2	5.1	4.7	4.6	4.4	3.2
No Odd										
Means	144.0	163.9	165.9	174.3	198.9	121.7	141.8	162.4	163.1	165.1
Medians	119.5	124.4	130.7	133.5	138.0	105.7	97.4	126.7	124.2	130.6
Skewness	2.1	1.7	1.4	2.5	3.2	1.8	1.8	2.2	1.8	1.4

3.2. Test 2: Short Run Forecast Comparison

Short term forecast is important as recent changes can be tracked because they can influence the following samples. The short-term forecast considers MA (with n-terms of size) and SES (with an alpha smoothing parameter), which were optimized to reduce forecasting errors. Both estimations have serious problems with non-stationary and trended time series, but under mild conditions they usually work very well in practice to smoothen data and filter the noise. Series with several outliers increase the complexity. In our case, to assure better results and to make a fair comparison, we optimize the respective parameters individually. The following section will discuss the parameter's optimization for SES and MA.

The optimal alpha was obtained for SES, minimizing the [weight/time] ratio based on the mean absolute percentage error (MAPE) estimator. The number of items to average for the rolling mean was also found by minimizing the MAPE estimator on the same ratio [weight/time], because it gives more relative importance to deviations. For extremely big outliers (infinite values), we consider the replacement with the median of the series, for fairness. The alpha estimated is near zero in almost all cases. The number of items to rolling average was between 2 and 3. These results seem natural for non-stationary series with high noise and many outliers. Based on the statistical test, it is not possible to assure that this estimator gives any better solution than the LR despite its complexity of tuning.

Even using these optimizations, the results did not improve, indicating that these methods by themselves do not give better results than LR because of the constant natural incertitude.

With the intention to properly evaluate the differences among the methods, we present Table 2 (for MAE) and Table 3. In these tables the results of the Sing Test, Wilcoxon Test and Two-Sample Test indicate SES and MA are significantly different than LR and Theil-Sen LR, but in the limit as they are not uniform for MAE and RMSE. LR and Theil-Sen LR behave similarly, being slightly significantly different for the RMS Odd (in the MAE case). The results for RMSE are slightly different possibly because the RMSE captures the influence of one outlier among the 4 points used to compare the forecasting, which encourages the use of the Theil-Sen LR. We should note the limitations of Wilcoxon Test and Two-sample Test with the data because these series are difficult to normalize (or at least to get in a symmetrical shape). Nevertheless, these tests must be performed to enlighten possible doubts from the Sign Test.

3.3. Discussion: Impacts and savings

Based on the results shown in the previous figures and tables, we notice that the biggest portion of the long queues is attributed to the cases with an Odd LR, and conversely, the queues are shorter for the No Odd LR cases. Moreover, based on the results shown in the Box-Cox figure, the Inter Quantile Range does not change considerably. Therefore, we can conclude that our Theil-Sen CR solves many of the LR practical problems, without jeopardizing punctual forecasting, but also that the local methods are not better than LR and Theil-Sen CR. Statistical tests shown are not conclusive to sustain any negative impact of Theil-Sen CR as the difference is of mixed significance for the sample size.

With the data in Table C.2, Table C.3 and Table C.4, it is possible to estimate the means of the current period and proposed periods. Considering the new

Table 2: Hypothesis Testing (Sign Test, Wilcoxon Test & Two Sample Test) for MAE

		Linear Regression		**Theil-Sen Constrained**		**Exponential Smoothing**		**Median Rates**		**Moving Average**	
MAE Odd (original)	**Sign Test**	H_0	Prob.	H_0	Prob.	H_0	Prob.	H_0	Prob.	H_0	Prob.
	Theil-Sen Constrained	False	0.058	—	—	**True**	0.022	False	0.401	False	0.141
	Linear Regression	—	—	False	0.058	**True**	—	False	0.177	**True**	0.005
(normalized)	**Wilcoxon Test**										
	Theil-Sen Constrained	False	0.052	—	—	**True**	—	False	0.147	**True**	0.005
	Linear Regression	—	—	False	0.052	**True**	—	**True**	0.027	**True**	—
(normalized)	**Two-sample Test**										
	Theil-Sen Constrained	False	0.523	—	—	**True**	0.013	False	0.768	False	0.058
	Linear Regression	—	—	False	0.523	**True**	0.002	False	0.361	**True**	0.014

		Linear Regression		**Theil-Sen Constrained**		**Exponential Smoothing**		**Median Rates**		**Moving Average**	
MAE No Odd (original)	**Sign Test**	H_0	Prob.	H_0	Prob.	H_0	Prob.	H_0	Prob.	H_0	Prob.
	Theil-Sen Constrained	False	0.058	—	—	**True**	0.022	False	0.401	False	0.141
	Linear Regression	—	—	False	0.058	**True**	—	False	0.177	**True**	0.005
(normalized)	**Wilcoxon Test**										
	Theil-Sen Constrained	False	0.052	—	—	**True**	—	False	0.147	**True**	0.005
	Linear Regression	—	—	**True**	0.027	**True**	—	**True**	0.027	**True**	—
(normalized)	**Two-sample Test**										
	Theil-Sen Constrained	False	0.523	—	—	**True**	0.013	False	0.768	False	0.058
	Linear Regression	—	—	False	0.361	**True**	0.002	False	0.361	**True**	0.014

In Sign Test, H_0 is the hypothesis that the difference has a distribution with zero median. In Wilcoxon, H_0 is the hypothesis that the difference has a (symmetrical) distribution with zero median. In Two-sample Test, H_0 is the hypothesis that both samples come from (independent) Gaussians with identical means. All hypothesis are rejected at $\alpha = 0.05$ of significance level.

Table 3: Hypothesis Testing (Sign Test, Wilcoxon Test & Two Sample Test) for RMS

		Linear Regression		**Theil-Sen Constrained**		**Exponential Smoothing**		**Median Rates**		**Moving Average**	
RMS Odd (original)	**Sign Test**	H_0	Prob.	H_0	Prob.	H_0	Prob.	H_0	Prob.	H_0	Prob.
	Theil-Sen Constrained	False	0.141	—	—	False	0.081	False	0.229	False	0.504
	Linear Regression	—	—	False	0.141	**True**	0.005	False	0.229	False	0.081
(normalized)	**Wilcoxon Test**										
	Theil-Sen Constrained	**True**	0.025	—	—	**True**	0.027	False	0.096	False	0.163
	Linear Regression	—	—	**True**	0.025	**True**	0.001	False	0.229	**True**	0.021
(normalized)	**Two-sample Test**										
	Theil-Sen Constrained	False	0.522	—	—	False	0.105	False	0.721	False	0.302
	Linear Regression	—	—	False	0.522	**True**	0.030	False	0.341	False	0.107

		Linear Regression		**Theil-Sen Constrained**		**Exponential Smoothing**		**Median Rates**		**Moving Average**	
RMS No Odd (original)	**Sign Test**	H_0	Prob.	H_0	Prob.	H_0	Prob.	H_0	Prob.	H_0	Prob.
	Theil-Sen Constrained	False	0.092	—	—	False	0.366	False	0.155	False	0.156
	Linear Regression	—	—	False	0.093	False	0.155	False	0.155	**True**	0.014
(normalized)	**Wilcoxon Test**										
	Theil-Sen Constrained	**True**	0.038	—	—	False	0.427	**True**	0.019	False	0.138
	Linear Regression	—	—	**True**	0.038	**True**	0.019	**True**	0.002	**True**	0.007
(normalized)	**Two-sample Test**										
	Theil-Sen Constrained	False	0.383	—	—	False	0.609	False	0.409	False	0.383
	Linear Regression	—	—	False	0.383	False	0.121	False	0.082	False	0.054

In Sign Test, H_0 is the hypothesis that the difference has a distribution with zero median. In Wilcoxon, H_0 is the hypothesis that the difference has a (symmetrical) distribution with zero median. In Two-sample Test, H_0 is the hypothesis that both samples come from (independent) Gaussians with identical means. All hypothesis are rejected at $\alpha = 0.05$ of significance level.

Table 4: Change in collecting periods

	(Rounded values)		
	Current Per.	**p-mean**	**% Change**
Mean	8	12	50.0%
Travels/year	46	31	32.6%

estimated average period ("p-mean"), it is also possible to estimate the reduction in the travels which indicates the possibility of enlarging at most in 50 % of the periods, as shown in Table 4, and to appreciate a 32.6 % reduction in the maximal number of travels. The reduction in the number of travels is important as fuel consumption and greenhouse gases (GHG) emissions are also reduced proportionally. It induces a better vehicle route planning that assigns less priority to those sites with lesser forecasted volume. Considering approximately 12.2 km between sites and a fleet of 200 trucks, a potential reduction of 940 tCO_2 equivalent per year might be achieved.

4. Conclusions

In this study, we propose a model that forecasts the daily input of industrial and commercial sites based on historical data. This is an interesting topic of research, as the recycling industry must deal with the problem of estimating the quantity of waste material generated by industrial and commercial sites that usually evolve in terms of production. Collecting operation decisions depends on management experience that is mostly based on imprecise information. As a result, several visits to the same client imply not only more cost for the recycling company but also more risk of spilling or product degradation in the other sites, and thus lower revenues and high environmental impacts. Moreover, as the number of collecting sites increases, the collection operations get more complicated, thus increasing costs and reducing revenues from recycled materials that otherwise would be landfilled.

Under the paradigms of the new economy and the development of smart cities, there is a temptation to consider complex technologies to support the planning of operations using geolocalization, measurement and communication integrated devices. New technology can increase communication costs and maintenance and still be prone to measuring errors. Real time information could not be enough for efficient operations, which need the selection and clustering of sites in advance to improve the truck routing efficiency, as variation in production can differ widely on time. The proposed forecasting method is suitable for helping the collection of several sites, based only on historical data.

Improving the routing operations for the collection of waste FOG can be done by using robust regressions, as proposed in this study. This regression provides unsupervised results when other estimations can get parameters not related to the process (like negative slopes in a filling tank). This regression gives slightly better results compared to SES and MA because of the variability of the process and the outliers. Short-term forecast methods usually need continuous recalculations, but the proposed Theil-Sen CR is very robust and it can be recalculated with new information, if needed. The main advantage of the proposed tool lies in the precision versus the simplicity of the automatic treatment of several series. This new estimator can be used as a trend indicator in complement with ARIMA, signaling the outliers that can be left outside the forecasting procedure.

As new venues for research, we suggest that current and lost clients could be studied first in their respective logistics routes to understand neighbors trends and then we could group homogeneous sites. The forecast of new sites could be improved by using the most recent information from neighbor sites having similar conditions (size, space, socio-economical population environment, franchise type, etc). The potential profitability of the new contracts could also be estimated using this association in combination with the vehicle routing costs. Other topics of

future research could be the use of dynamical models and the use of data mining for clustering sites and find evidences of commodity correlations.

Acknowledgements

Funding for this research has been provided by CRSNG/NSERC (Natural Sciences and Engineering Research Council of Canada). In particular, we are grateful to our industrial partner (Sanimax Inc.). We would like to thank Robert Gregoire, Maxime Gil-Blaquiere, Amine Chbicheb, Jean-Francois Paquette and Patrick Brodeur for their valuable comments and suggestions for this project.

5. References

Aksen, D., Kaya, O., Salman, F.S., Akça, Y., 2012. Selective and periodic inventory routing problem for waste vegetable oil collection. Optimization Letters 6, 1063–1080.

Andrews, A., Gregoire, M., Rasmussen, H., Witowich, G., 2013. Comparison of recycling outcomes in three types of recycling collection units. Waste Manag. 33, 530–535.

Arebey, M., Hannan, M.A., Basri, H., Begum, R.A., Abdullah, H., Arebey, M., Hannan, M.A., Basri, H., Begum, R.A., Abdullah, H., 2010. Solid waste monitoring system integration based on RFID, GPS and camera, in: 2010 International Conference on Intelligent and Advanced Systems, ICIAS 2010, IEEE. pp. 1–5.

Arena, U., Di Gregorio, F., 2014. A waste management planning based on substance flow analysis. Resources, Conservation and Recycling 85, 54–66.

Bagnall, A., Davis, L., Hills, J., Lines, J., 2012. Transformation Based Ensembles for Time Series Classification. Proceedings of the 2012 SIAM International Conference on Data Mining , 307–318.

Bialy, H.E., Gomaa, O.M., Azab, K.S., 2011. Conversion of oil waste to valuable fatty acids using Oleaginous yeast. World Journal of Microbiology and Biotechnology 27, 2791–2798.

Campbell, A.M., Wilson, J.H., 2014. Forty years of periodic vehicle routing. Networks 63, 2–15.

Faria, J., Sousa, A., Reis, A., Filipe, V., Barroso, J., 2016. Probe and sensors development for level measurement of fats, oils and grease in grease boxes. Sensors (Switzerland) 16, 1517.

Fitzgerald, G.C., Krones, J.S., Themelis, N.J., 2012. Greenhouse gas impact of dual stream and single stream collection and separation of recyclables. Resour. Conserv. Recycl. 69, 50–56.

Folianto, F., Low, Y.S., Yeow, W.L., 2015. Smartbin: Smart waste management system, in: 2015 IEEE 10th International Conference on Intelligent Sensors, Sensor Networks and Information Processing, ISSNIP 2015, pp. 1–2.

Guabiroba, R.C.d.S., Silva, R.M.d., César, A.d.S., Silva, M.A.V.d., 2017. Value chain analysis of waste cooking oil for biodiesel production: Study case of one oil collection company in Rio de Janeiro - Brazil. Journal of Cleaner Production 142, 3928–3937.

Gutierrez, J.M., Jensen, M., Henius, M., Riaz, T., 2015. Smart Waste Collection System Based on Location Intelligence. Procedia Computer Science 61, 120–127.

Han, H., Ponce-Cueto, E., Cueto, E.P., 2015. Waste Collection Vehicle Routing Problem: A Literature Review. PROMET - Traffic&Transportation 27, 345–358.

Hancke, G.P., de Silva, B.d.C., Hancke, G.P., 2013. The role of advanced sensing in smart cities. Sensors (Switzerland) 13, 393–425.

Hannan, M.A., Abdulla Al Mamun, M., Hussain, A., Basri, H., Begum, R.A., 2015. A review on technologies and their usage in solid waste monitoring and management systems: Issues and challenges. Waste management (New York, N.Y.) 43, 509–23.

Hirsch, R.M., Slack, J.R., Smith, R.A., 1982. Techniques of trend analysis for monthly water quality data. Water Resources Research 18, 107–121.

Huddleston, S.H., Porter, J.H., Brown, D.E., 2015. Improving forecasts for noisy geographic time series. Journal of Business Research 68, 1810–1818.

Hyndman, R.J., Khandakar, Y., 2008. Automatic time series forecasting: The forecast package for R. Journal Of Statistical Software 27, C3–C3.

Lam, S.S., Wan Mahari, W.A., Jusoh, A., Chong, C.T., Lee, C.L., Chase, H.A., 2017. Pyrolysis using microwave absorbents as reaction bed: An improved approach to transform used frying oil into biofuel product with desirable properties. Journal of Cleaner Production 147, 263–272.

Laurent, A., Bakas, I., Clavreul, J., Bernstad, A., Niero, M., Gentil, E., Hauschild, M.Z., Christensen, T.H., 2014. Review of LCA studies of solid waste management systems - Part I: Lessons learned and perspectives. Waste Management 34, 573–588.

Mes, M., 2012. Using Simulation to Assess the Opportunities of Dynamic Waste Collection in Use Cases of Discrete Event Simulation: Appliance and Research. Springer Berlin Heidelberg, Berlin, Heidelberg. chapter 13. pp. 277–307. Editor: Bangsow, Steffen.

Mes, M., Schutten, M., Rivera, A.P., 2014. Inventory routing for dynamic waste collection. Waste management (New York, N.Y.) 34, 1564–76.

Nadeem, H., Habib, N.Z., Ng, C.A., Zoorob, S.E., Mustaffa, Z., Chee, S.Y., Younas, M., 2017. Utilization of catalyzed waste vegetable oil as a binder for the production of environmentally

friendly roofing tiles. Journal of Cleaner Production 145, 250–261.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al., 2011. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, 2825–2830.

Ramos, T.R.P., Gomes, M.I., Barbosa-Póvoa, A.P., 2013. Planning waste cooking oil collection systems. Waste management (New York, N.Y.) 33, 1691–703.

Sen, P.K., 1968. Estimates of the Regression Coefficient Based on Kendall's Tau. Journal of the American Statistical Association 63, 1379–1389.

Sharma, V.K., Chandna, P., Bhardwaj, A., 2017. Green supply chain management related performance indicators in agro industry: A review. Journal of Cleaner Production 141, 1194–1208.

Singh, H., Garg, R.K., Sachdeva, A., 2018. Supply chain collaboration: A state-of-the-art literature review. Uncertain Supply Chain Management 6, 149–180.

Tarr, G., Müller, S., Weber, N.C., 2012. A robust scale estimator based on pairwise means. Journal of Nonparametric Statistics 24, 187–199.

Taylor, S.J., Letham, B., Taylor, S.J., Letham, B., 2017. Forecasting at Scale. Unpublished Work , 1–17.

Team, R.D.C., 2004. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing 739.

Teixeira, J., Antunes, A.P., de Sousa, J.P., 2004. Recyclable waste collection planning—a case study. European Journal of Operational Research 158, 543–554.

Tonjes, D.J., Mallikarjun, S., 2013. Cost effectiveness of recycling: A systems model. Waste Management 33, 2548–2556.

Tran, N., Tran, C., Ho, P., Hall, P., McMurchie, E., Hessel, V., Ngothai, Y., 2016. Extraction of fats, oil and grease from grease trap waste for biodiesel production. Renewable Energy .

Tu, Q., McDonnell, B.E., 2016. Monte Carlo analysis of life cycle energy consumption and greenhouse gas (GHG) emission for biodiesel production from trap grease. Journal of Cleaner Production 112, 2674–2683.

Vinyes, E., Oliver-Solà, J., Ugaya, C., Rieradevall, J., Gasol, C.M., 2013. Application of LCSA to used cooking oil waste management. International Journal of Life Cycle Assessment 18, 445–455.

Wilcox, R.R., 1998. A note on the Theil-Sen regression estimator when the regressor is random and the error term is heteroscedastic. Biometrical Journal 40, 261–268.

Yan, W., 2012. Toward automatic time-series forecasting using neural networks. IEEE Transactions on Neural Networks and Learning Systems 23, 1028–1039.

Zsigraiova, Z., Semiao, V., Beijoco, F., 2013. Operation costs and pollutant emissions reduction by definition of new collection scheduling and optimization of MSW collection routes using GIS. The case study of Barreiro, Portugal. Waste Manag. 33, 793–806.

- We propose a model that forecasts the daily oil waste of thousands of industrial and commercial sites of the agro-food industry based on historical data.
- The model is based on the Theil-Sen regression and rejects errors and outliers to simplify the forecast of future collections.
- We apply this method to a real case study and compare its performance at different collecting sites.
- The results, based on the average error, achieve a maximum potential of 32.6% fewer visits by enlarging periodical visits until reaching 50.0% of the current period.