More than one Author with different Affiliations

Jerónimo Hernández-González*[1], Iñaki Inza[1], Igor Granado[2], Oihane C. Basurko[2], Jose A. Fernandes[2] and Jose A. Lozano[1,3]

[1]Department of Computer Science and Artificial Intelligence, University of the Basque Country UPV/EHU, Donostia, Spain.
[2]Marine Research Division at AZTI-Tecnalia, Pasaia, Spain.
[3]Basque Center for Applied Mathematics, Bilbao, Spain.

Abstract

In regression, a predictive model which is able to anticipate the output of a new case is learnt from a set of previous examples. The output or response value of these examples used for model training is known. When learning with aggregated outputs, the examples available for model training are individually unlabeled. Collectively, the aggregated outputs of different subsets of training examples are provided. In this paper, we propose an iterative methodology to learn linear models from this type of data. In spite of being simple, its competitive performance is shown in comparison with a straightforward solution and state-of-the-art techniques. A real world problem is also illustrated which naturally fits the aggregated outputs framework: the estimation of marine litter beaching along the south-east coastline of the Bay of Biscay.

Index terms— Machine learning, Regression, Linear models, Aggregated outputs, Expectation-Maximization, Marine litter beaching

1 Introduction

In supervised learning, a model is learnt from a set of previous examples of the problem of interest in order to predict, for new unseen examples, the value of the response or output variable. When the response variable is categorical, the framework is known as classification, whereas it is named as regression when the output is continuous. In both frameworks, the term *supervised* learning indicates that all the examples used for model training are provided together with their real output value, a.k.a. ground truth. However, obtaining the ground truth is usually hard and costly and normally the supervision of the training examples

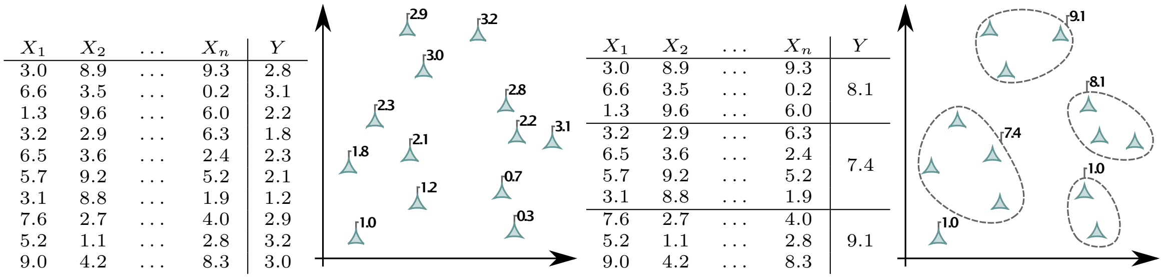
*jeronimo.hernandez@ehu.eus

is not complete. Many methodologies have been proposed to learn from these partially or weakly supervised datasets [19].

In this paper, we focus on the aggregated outputs (AO) framework [27], a weakly supervised regression problem. The characteristic training dataset of this framework is divided into disjoint subsets of examples and the individual real output value of each training example is not provided. Alternatively, for each subset of examples, a single output value is available: the total output value aggregated from all the examples of the subset. Musicant et al. [27] approached the problem for the first time with the adaptation of three classical regression techniques: support vector regression (SVR), k-nearest neighbor (KNN) and artificial neural network (ANN). Afterward, it received little attention in comparison with learning from label proportions [21, 36, 5, 38, 43, 18, 31], its equivalent framework in classification. Here, the response values of a subset of examples are aggregated in the form of proportions of examples belonging to each label. The related literature covers support vector machines [38, 47, 35], discriminant analysis [31], boosting [34], clustering-based approaches [5, 43] or Expectation-Maximization based probabilistic approaches [18]. Theoretical guarantees have also been provided [30, 11]. These methods have been applied to real domains such as spam filtering [36], poll prediction [31, 44], embryo selection [17], fraud detection [38], manufacturing [43], brain-computer interfaces [20], high energy physics [8], etc. In regression, on the contrary, this genuine aggregated response has only been described, so far, for estimating the amount of black carbon in aerosol particles [27].

The problem of marine litter beaching estimation, from the area of environmental sciences, fits the AO framework as well. Marine litter is defined as any persistent, manufactured or processed material abandoned or disposed of in the marine environment [14]. In recent times, concern about its ecological, social and economic impact [7, 42, 26, 15] has grown. Beach litter has a negative impact on marine ecosystems [9, 15], with well-known issues such as ghost fishing [22, 23, 25] or plastic ingestion [1, 24, 29, 41]; international institutions [12] have warned that by 2050 oceans might contain more plastic than fish. Beach litter also harms the recreational and aesthetic value of coastal areas, affects public health [16, 3] and impacts on industrial sectors such as fisheries, shipping or tourism [28, 16]. Accumulated beach litter is mostly related with (i) holiday tourism [32], (ii) waste unburied by spring tides [2], (iii) fishing apparatus [46], and (iv) organic material or inland produced litter deposited by rivers and tides [40]. Beach cleaning services spend large amounts of money annually on removing this waste. Cleaning on a daily basis is unfeasible due to its high economic cost. Alternatively, cleaning services need to plan and set up resource allocation and, to do so, an estimation of the accumulation of litter is necessary.

Our aim here is to learn a model that daily predicts the accumulation of waste on a beach based on environmental conditions. To describe an example (a day), metocean and environmental variables are used [33, 37]. The output variable represents the amount of litter accumulated on a beach during 24 hours. Metocean and environmental observations are collected everyday by automatic stations. However, the amount of waste is only measured when service members



X_1	X_2	...	X_n	Y
3.0	8.9	...	9.3	2.8
6.6	3.5	...	0.2	3.1
1.3	9.6	...	6.0	2.2
3.2	2.9	...	6.3	1.8
6.5	3.6	...	2.4	2.3
5.7	9.2	...	5.2	2.1
3.1	8.8	...	1.9	1.2
7.6	2.7	...	4.0	2.9
5.2	1.1	...	2.8	3.2
9.0	4.2	...	8.3	3.0

X_1	X_2	...	X_n	Y
3.0	8.9	...	9.3	8.1
6.6	3.5	...	0.2	
1.3	9.6	...	6.0	
3.2	2.9	...	6.3	7.4
6.5	3.6	...	2.4	
5.7	9.2	...	5.2	
3.1	8.8	...	1.9	
7.6	2.7	...	4.0	9.1
5.2	1.1	...	2.8	
9.0	4.2	...	8.3	

Figure 1: Simulated example of datasets of regular regression (left) and aggregated outputs (right), together with the respective graphical descriptions.

personally visit the beach to remove it. In this way, the training data consists of a set of examples, each of which describes the coastal line of a municipality on a specific day by means of a *complete* vector of metocean and environmental measurements, and an output value which might be *missing*. Moreover, on a day without cleaning service, waste is neither measured nor collected and, thus, it accumulates for the following day. Therefore, when a cleaning team visits the beach, the accumulated waste is an aggregated measurement for the period starting right after the previous removal, possibly covering a few days (a subset of training examples).

The main contributions of this paper are two-fold:

- A novel strategy to infer regression models from a dataset with aggregated outputs and its illustration with linear models.
- A novel approach to marine litter beaching prediction, with a case study on real data from the Basque coastline.

The use of linear models offers a simple, yet efficient, solution with a closed-form expression for its learning procedure. Nevertheless, the underlying strategy could be straightforwardly adapted to learn other types of (non-)linear models. An extensive empirical study shows that the approach presented is competitive with respect to previous proposals in both synthetic and real data. This estimation of marine litter beaching may provide the environmental research community with an alternative predictive approach, which makes the most of the available information for model training.

The paper continues as follows. Firstly, after the presentation of background concepts and techniques, a formal description of the AO problem is given and our method for inferring linear models is presented. Its performance is tested and compared on synthetic AO data with state-of-the-art learning methods. Next, the case study is presented and the previously considered methods are applied. Finally, after the discussion of the experimental results, open questions are put forward and conclusions are drawn.

2 Background

In this paper, we deal with the AO framework by means of linear models. Linear models are a typical solution to regression problems. In the standard completely supervised framework, where each training example is individually supervised and standard supervised learning techniques can be used, linear models are mainly learned by means of the method of least squares.

2.1 Regression

A supervised learning problem is described by a set of n explanatory features and a special feature, the response or output variable. Specifically, the response value of a regression problem is continuous. In a regression problem, the objective is to train, using a set of examples, a regression model which anticipates the response value of new examples. A problem example $(\boldsymbol{x}, y)$ is a $(n+1)$-tuple where each feature takes a specific value. For model training, a dataset of m *fully* supervised examples, $\{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_m, y_m)\}$, which are supposed to be independently and identically distributed (i.i.d.) samples from some underlying probability distribution, is provided. An example is fully supervised if the value of the response variable y is known and not missing.

The training data is sometimes represented in matrix form by means of a $(m \times n)$-matrix X and a (m)-vector $\boldsymbol{y}$. The j-th row of matrix X represents the j-th example of the training data, $\boldsymbol{x}_j$, whereas the corresponding response value, y_j is the j-th entry of vector $\boldsymbol{y}$. The entry x_{jv} in matrix X represents the value of the v-th explanatory variable for the j-th example.

2.2 Linear models

The objective of linear regression is to infer a linear model (LM) that approaches the relationship between the explanatory variables and the response. The existence of a vector of parameters, $\boldsymbol{\beta}$, such that the response variable is a linear function of the explanatory variables, is assumed,

$$y_j = \boldsymbol{x}_j^t \boldsymbol{\beta} + \epsilon_j, \quad \forall j \in \{1, \ldots, m\}$$

where ϵ_j are the regression residuals. In matrix form, the problem can be stated as,

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $\boldsymbol{\epsilon}$ is the vector of residuals. Generally, in practice, the original explanatory vectors $\boldsymbol{x}_j$ (and also the matrix X) are enlarged with a constant value $(1, \boldsymbol{x}_j)$. This imposes the use of an extra parameter, β_0, called the intercept, which guarantees that the sum of the residuals ϵ_j is zero. In this standard supervised learning framework, the vector of responses $\boldsymbol{y}$ is assumed to be complete (fully supervised examples).

Least squares is the most common method to train a linear model for multiple regression. Let us define the regression residuals considering that a (linear) model might not fit every single data point. Thus, a residual ϵ_j is defined as the difference between the *estimated* response value, $\boldsymbol{x}_j^t\boldsymbol{\beta}$, and the *real* response value, y_j,

$$\epsilon_j = y_j - \boldsymbol{x}_j^t\boldsymbol{\beta}$$

for a given fit $\boldsymbol{\beta}$ of the model. Note that a measure of the error of a model fit $\boldsymbol{\beta}$ can be obtained as the sum of squared residuals,

$$s(\boldsymbol{\beta}) = \sum_{j=1}^{m}(y_j - \boldsymbol{x}_j^t\boldsymbol{\beta})^2 \tag{1}$$

The least squares method produces the set of parameters $\hat{\boldsymbol{\beta}}$ that minimizes the sum of squared residuals (Equation 1),

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^{(n+1)}} s(\boldsymbol{\beta}) \tag{2}$$

which, in matrix form, has the following closed-form expression,

$$\hat{\boldsymbol{\beta}} = (X^t \cdot X)^{-1} \cdot X^t \cdot \boldsymbol{y} \tag{3}$$

where $\boldsymbol{y}$ is the original response vector and $(X^t \cdot X)^{-1} \cdot X^t$ is the pseudo-inverse of matrix X. When it is used for prediction, given a new explanatory vector, $\boldsymbol{x}$, an estimation of its response value is obtained as,

$$\hat{y} = \boldsymbol{x}^t\hat{\boldsymbol{\beta}} \tag{4}$$

3 Aggregated Outputs

The main novelty of the learning with aggregated outputs framework is the lack of a fully supervised set of training examples (see Figure 1). The m explanatory vectors are individually unsupervised (without associated response value) and grouped into b *bags*, $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m\} = \boldsymbol{B}_1 \cup \boldsymbol{B}_2 \cup \cdots \cup \boldsymbol{B}_b$, where $\boldsymbol{B}_i \cap \boldsymbol{B}_{i'} = \emptyset, \forall i \neq i'$. Each bag $\boldsymbol{B}_i = \{\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, \ldots, \boldsymbol{x}_{im_i}\}$ groups m_i instances, with $\sum_{i=1}^{b} m_i = m$. Together with each bag, a limited piece of supervision is provided: the aggregated output $\bar{y}_i$ is the sum of the individual response values of all the examples in $\boldsymbol{B}_i$ ($\bar{y}_i = \sum_{j=1}^{m_i} y_{ij}$, where the individual values y_{ij} are missing). In matrix form, the descritive $(m \times n)$-matrix X is completely available, in contrast with the instance-wise (length m) response vector $\boldsymbol{y}$, which is missing. Alternatively, a bag-wise (length b) vector, $\bar{\boldsymbol{y}}$, of aggregated responses is available, where the i-th entry represents the aggregated output of $\boldsymbol{B}_i$. The rest of the framework is defined similarly to the standard regression framework previously presented. Among these similarities, the most important one is that the objective of the AO framework is also to learn a regression model that predicts the response value of new unseen examples.

An aggregated output involves uncertainty in the measurement of the output variable. Specifically, the level of uncertainty depends on the characteristics of the corresponding bag. Intuitively, the larger the bag, the more possible ways there are to distribute the aggregated response, $\bar{y}_i$, among the m_i individual responses, y_{ij}, and, thus, the higher the degree of uncertainty in the response of the individual examples of the bag. On the contrary, the smaller the number of examples in the bag, m_i, the larger the certainty. In the extreme, a bag with a single example ($m_i = 1$) involves full certainty as $y_{i1} = \bar{y}_i$. In problems with large bags, therefore, the performance of the learning techniques is expected to be compromised.

3.1 Linear models from aggregated outputs

Although the learning procedure is necessarily different, the linear models learnt with AO are essentially similar to those learnt with complete supervision. As the instance-wise response vector, $\boldsymbol{y}$, is not complete, ordinary least squares cannot be applied as explained in the previous section. The bag-wise vector of aggregated outputs, $\bar{\boldsymbol{y}}$, is the available information of supervision. At this initial point, in order to apply the aforementioned techniques for training linear models, one possibility is to transform the vector of aggregated outputs, $\bar{\boldsymbol{y}}$, into an instance-wise vector, $\boldsymbol{y}'$, that assigns a certain response to each example.

A naive transformation of the aggregated response vector, $\bar{\boldsymbol{y}}$, is to equally distribute (disaggregate) the aggregated response, $\bar{y}_i$, among all the instances of bag $\boldsymbol{B}_i$, $y'_{ij} = \bar{y}_i/m_i, \forall i, j : i \in \{1, \ldots, b\} \wedge j \in \{1, \ldots, m_i\}$. In matrix form, it can be defined as,

$$\boldsymbol{y}' = A^t(\bar{\boldsymbol{y}} \oslash (A\boldsymbol{o})) \tag{5}$$

where A is the assignment $(b \times m)$-matrix —it codifies the assignment of examples to bags: $A_{ij} = 1$ if example $\boldsymbol{x}_j$ belongs to bag $\boldsymbol{B}_i$, and $A_{ij} = 0$ otherwise—, $\boldsymbol{o}$ is an m-tuple with all its entries $o_j = 1$, and $\oslash$ represents the Hadamard division or entry-wise division of two equal-sized vectors ($\boldsymbol{a} = \boldsymbol{b} \oslash \boldsymbol{c} \equiv a_i = b_i/c_i, \forall i$). As vector $\boldsymbol{y}'$ is instance-wise, it can be used together with matrix X to infer the linear model parameters, $\boldsymbol{\beta}'$, by means of Equation 3. This procedure provides a first fit of a linear model. However, this fit is likely to be deficient as the response vector $\boldsymbol{y}'$, a naively disaggregated instance-wise output, is used for its training.

Inspired by the Expectation-Maximization strategy [6], the proposed learning approach is an iterative methodology of two steps. Firstly, given an instance-wise response vector $\boldsymbol{y}'$ (disaggregated from the aggregated outputs $\bar{\boldsymbol{y}}$) and matrix X, a linear model is inferred. In turn, the current fit of the model is used to improve the disaggregation of $\bar{\boldsymbol{y}}$ to obtain a new estimate of $\boldsymbol{y}'$. From an alternative point of view, the method reduces to the iterative re-estimation of two instance-wise response vectors: (i) an unconstrained prediction of the current fit of the linear model for all the training examples,

$$\boldsymbol{y}'' = X\ \boldsymbol{\beta}' = X\ (X^t\ X)^{-1}\ X^t\ \boldsymbol{y}' \tag{6}$$

and, (ii) a transformation of the current estimate $\boldsymbol{y}''$ to guarantee that the instance-wise outputs sum up to the bag-wise aggregated outputs, $\bar{\boldsymbol{y}}$,

$$\boldsymbol{y}' = A^t \; (\bar{\boldsymbol{y}} \oslash (A \; \boldsymbol{y}'')) \odot \boldsymbol{y}'' \tag{7}$$

where $\odot$ represents the Hadamard product or entry-wise product of two equal-sized vectors ($\boldsymbol{a} = \boldsymbol{b} \odot \boldsymbol{c} \equiv a_i = b_i \cdot c_i, \forall i$).

In other words, in Equation 6 a fit $\boldsymbol{\beta}'$ of the linear model is learnt given $\boldsymbol{y}'$ (by Equation 3) and used to obtain an estimation of the responses of each training example $\boldsymbol{y}''$ (by Equation 4). The produced estimation is not guaranteed to fulfill the only available information of supervision: the aggregated outputs. Taking full advantage of the information of supervision, Equation 7 transforms $\boldsymbol{y}''$ into a proportional estimate, $\boldsymbol{y}'$, that does concur with the aggregated outputs, $\bar{\boldsymbol{y}}$. This new instance-wise estimation, $\boldsymbol{y}'$, can be used to feed Equation 6 again.

To sum up, our proposal obtains an initial estimation of $\boldsymbol{y}'$ by Equation 5 and, subsequently, iterates Equations 6 and 7 until the minimum $\hat{\boldsymbol{\beta}}^*$ parameters that lead to the best results are reached. The calculation of the model parameters, $\hat{\boldsymbol{\beta}}$, is an implicit sub-step of Equation 6.

This iterative technique is empirically compared in the following sections with straightforward approaches and with state-of-the-art methods presented by Musicant et al. [27].

3.2 Experimental design

The lack of publicly available real AO datasets makes any attempt to test a novel technique hard. In this first part of the paper, we test our proposal in synthetically aggregated outputs obtained from the transformation of fully supervised regression data. Specifically, up to 16 regular regression datasets were collected from three public repositories [4, 45, 13] (see Table 1). The strategy of Musicant et al. [27] was followed to transform a fully supervised dataset into an AO dataset: (i) examples are ordered by increasing order of their response value ($\{(\boldsymbol{x}_j, y_j)\}_{j=1}^m : y_j > y_{j'} \rightarrow j > j'$) (ii) a certain rate r of pairs of examples are randomly selected, (iii) for each pair of examples, their positions in the ordering are swapped, (iv) groups (bags) of m_i consecutive examples in the ordering are formed, and (v) the individual responses are aggregated by bag ($\{y_j\}_{j=1}^m$ into $\{\bar{y}_i\}_{i=1}^b$), and then removed. It is worth noting that the order confers some sort of information of supervision. In order to control the experimental settings, two parameters allow us to simulate scenarios of different complexity: the bag size, m_i, and the swap rate, r, defined as a proportion of the total size of the dataset. For example, a value $r = 0.5$ implies that, after ordering the dataset, $0.5 \cdot m$ (half the size of the dataset) randomly selected pairs of examples are swapped. By shuffling *larger proportions* of examples before the division into *larger bags*, a higher degree of uncertainty is induced. Small values of both m_i and r parameters tend to produce AO datasets with more certain information of supervision.

Model validation with only weakly supervised data is an open issue. In this paper, as aggregated outputs are synthetically aggregated, we do have real class labels for model validation. The original dataset is first divided into training and validation data. The partition of the data devoted for model training is transformed following the aforementioned procedure, whereas the partition used for model validation remains untouched. All the experiments are evaluated by means of 10 × 5-fold cross validation.

Our experimental setting is two-fold. Firstly, in order to test the usefulness of the aggregated outputs, our proposal is compared with linear models learnt in the fully supervised scenario and a control approach. Secondly, our proposal is compared with other methods previously presented in the related literature. The numerical results of all these experiments are publicly available in the website associated with this study[1].

3.3 Assessing the contribution of the aggregated outputs to learn linear models

In this first set of experiments, we aim to show the benefits of using a learning technique specifically designed to learn with aggregated outputs. Our AO solution is compared with two baselines: (i) a fully supervised approach (linear regression models learnt with the original untransformed response variable), and (ii) a control approach (linear models learnt with a response vector naively disaggregated using Equation 5).

In these experiments, different scenarios of aggregated outputs are designed by the use of up to 16 different datasets, four different bag sizes, $m_i = \{3, 5, 10, 20\}$, and seven different numbers of swapped instance pairs, $r = \{0, 0.05, 0.1, 0.2, 0.4, 1, 2\}$. A detailed collection of the experimental results is shown in Figure 2, where performance is measured by means of root mean square error (RMSE), with each each subfigure showing a different y-axis (error) range. Additionally, the averaged experimental results with all the datasets are summarized in Figure 3. Due to the large divergences among the error ranges of the different datasets, this figure shows the average error relative to the result of the linear model learnt with full supervision.

According to the results in Figure 2, the linear models learnt with AO out-

[1] http://www.sc.ehu.es/ccwbayes/members/jeronimo/aobeaches/

Table 1: Real datasets used [4, 45, 13] described by number of examples (m) and number of explanatory variables (n).

Dataset	m	n	Dataset	m	n
machine	209	6	mg	1385	6
bodyfat	252	14	airfoil	1503	6
eunite2001	336	16	space_ga	3107	6
mpg	392	7	abalone	4177	8
boston_housing	506	13	winequality	6497	12
stock	950	9	kinematics	8192	8
swd	1000	11	cpusmall	8192	12
concrete	1030	9	bank	8192	32

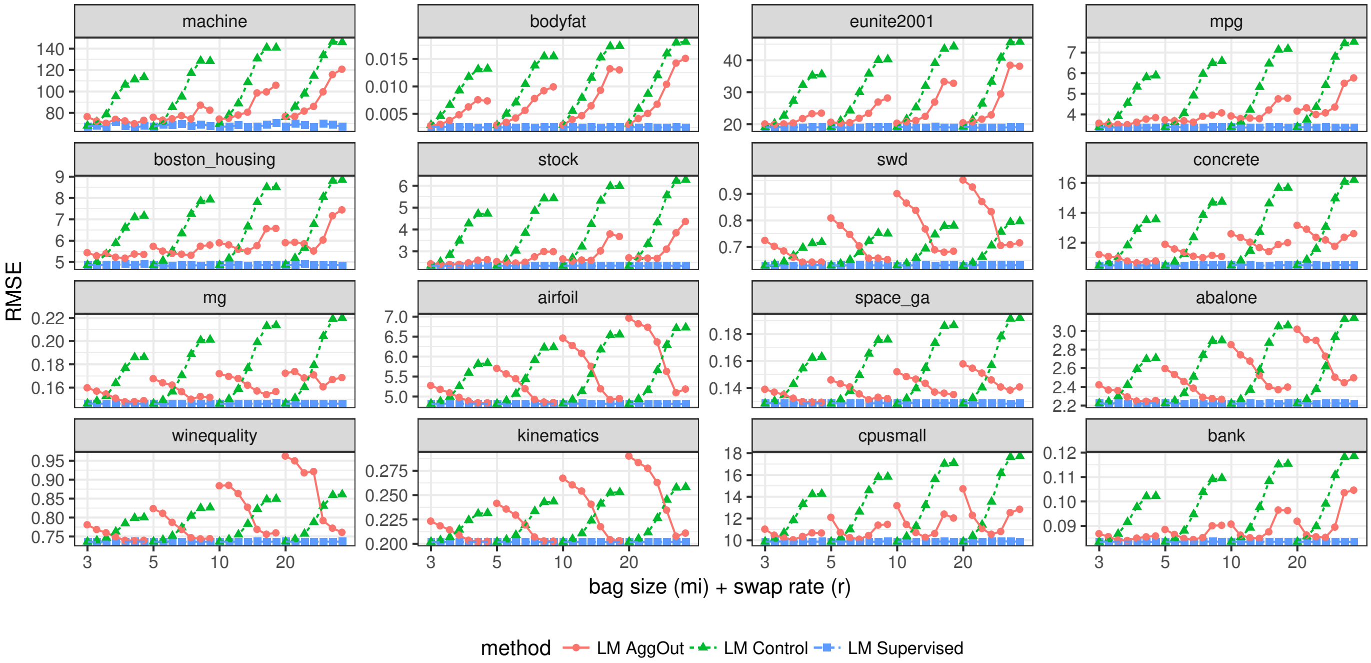


Figure 2: Results in terms of root mean square error of linear models learnt with aggregated outputs and with equally distributed outputs (control) in different simulated AO scenarios, and also with full supervision in the original scenario. For generating synthetic AO scenarios, different bag sizes, $m_i = \{3, 5, 10, 20\}$ and swap rates, $r = \{0, 0.05, 0.1, 0.2, 0.4, 1, 2\}$, are used. Each subfigure shows the results of the experiments in a different dataset.

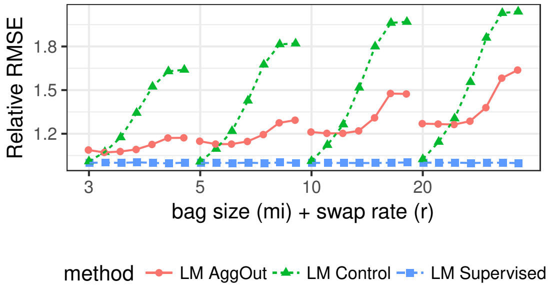


Figure 3: Results in terms of average *relative* root mean square error of models learnt with aggregated outputs and with the control approach. Results are relative to those of a linear model learnt with full supervision and averaged over all the datasets (Table 1). Different AO scenarios are simulated using different bag sizes, $m_i = \{3, 5, 10, 20\}$ and swap rates, $r = \{0, 0.05, 0.1, 0.2, 0.4, 1, 2\}$.

perform those learnt with the control approach when the number of swaps performed during the AO simulation procedure increases. Note that the first point in each line of the figure represents an experiment carried out without swaps, that is, the examples are ordered by response value before bag aggregation. In this scenario, by definition, the first bag has the smallest aggregated response value and, incrementally, the last bag has the largest. This configuration, which guarantees the lowest variance in the bags, brings determinant information of supervision: the control version consistently performs as accurately as the fully supervised approach. Moreover, with large datasets (second column, Table 1), there always exist aggregated scenarios where the AO models are competitive regarding the fully supervised approach. This is an indicator that, also in the AO framework, a larger number of examples enhances the performance of the learnt models. Finally, the bag size, m_i, seems to affect both the control and the AO approaches in a similar way. In most of the subfigures of Figure 2, the plots for the different bag sizes are similar, with proportionally enlarged differences.

It is worth noting that the aforementioned behaviors are extreme in the case of large datasets ($m > 1000$), where the control and the AO approaches show opposite behaviors. The control approach (notably) outperforms the AO solution when few swaps are used for AO simulation, and, with a large number of swaps, it is the other way around. With *cpusmall* and *bank* domains, the largest datasets in terms of number of samples (m) and explanatory variables (n), the control approach is only competitive with no or few swaps. These results are in line with the idea that the number of explanatory variables determines the sample size required to learn a robust model.

The summarized relative results in Figure 3 verify the previously exposed behaviors. The bag size, m_i, affects both approaches (AO and control) similarly. As the swap rate, r, is enlarged, the performance of both approaches differs. With no swaps, the control approach stands out and, with very small swap rates ($r \sim 0.05$), it is still competitive with (although not better than) the AO approach. With a considerable number of swaps, the AO solution systematically outperforms the control approach.

3.4 Comparison with state-of-the-art

Once the benefits of using a specifically designed AO strategy have been shown, our proposal is compared with state-of-the-art strategies. All three methods developed by Musicant et al. [27] (SVR, KNN and ANN) were tested, although the results of ANN models are not shown in this paper as they were consistently worse than those of the rest of techniques. Their inclusion would hardly provide any valuable insight and, nevertheless, would make the appreciation of differences among the rest of the techniques difficult.

Similar to the previous experiments, different AO scenarios are designed by the use of the 16 datasets, four different bag sizes, $m_i = \{3, 5, 10, 20\}$, and seven different swap rates, $r = \{0, 0.05, 0.1, 0.2, 0.4, 1, 2\}$. The results in terms of RMSE are displayed in Figure 4 for each dataset separately. Additionally, Figure 5 summarizes the results over all the datasets by means of the average

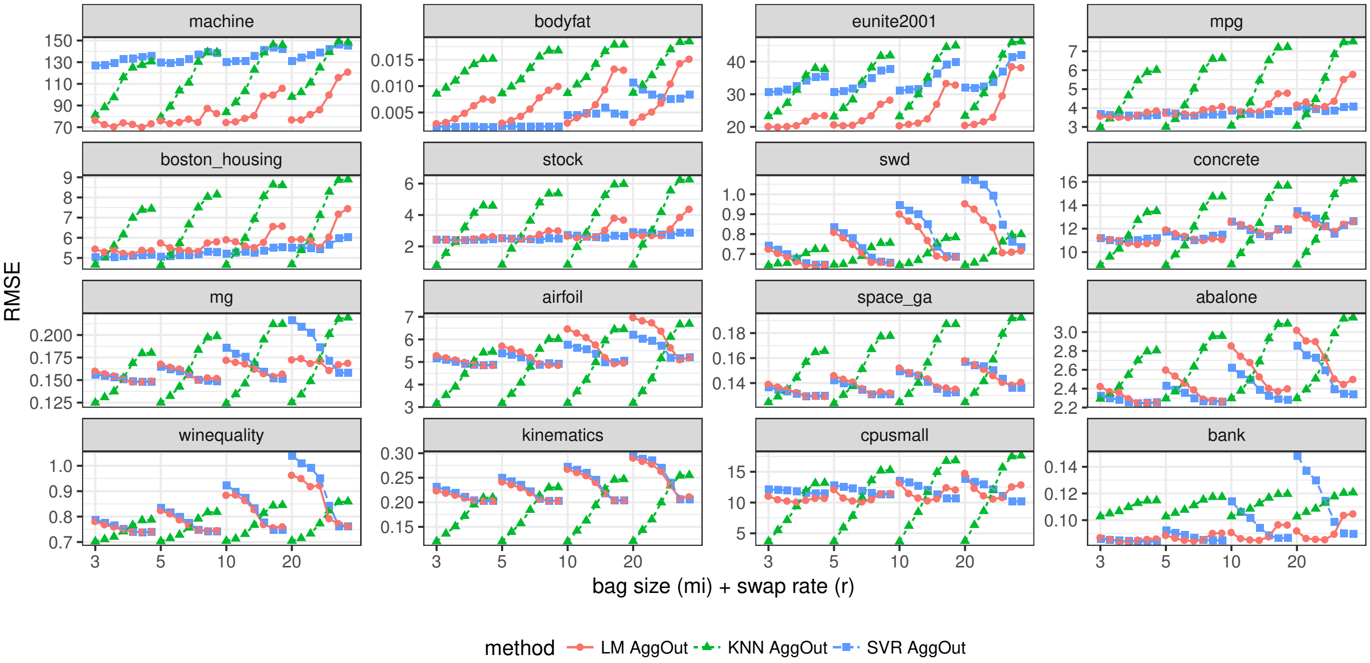


Figure 4: Results in terms of root mean square error of linear, KNN and SVR models learnt with aggregated outputs, in different simulated AO scenarios. Different bag sizes, $m_i = \{3, 5, 10, 20\}$ and swap rates, $r = \{0, 0.05, 0.1, 0.2, 0.4, 1, 2\}$, are used for generating synthetic AO scenarios. Each subfigure shows the results of the experiments in a different dataset.

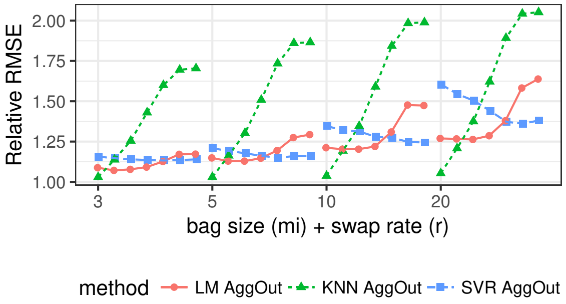


Figure 5: Results in terms of average *relative* root mean square error of linear, KNN and SVR models learnt with aggregated outputs. Results are relative to those of a linear model learnt with full supervision and averaged over all the datasets (Table 1). Different AO scenarios are simulated using different bag sizes, $m_i = \{3, 5, 10, 20\}$ and swap rates, $r = \{0, 0.05, 0.1, 0.2, 0.4, 1, 2\}$.

relative error. As aforementioned, the relative error is defined as the error obtained by an AO technique divided by that of a linear model learnt in the fully supervised scenario.

The experimental results displayed in Figure 4 show the competitive behavior of our proposal based on linear models learnt from AO. In several datasets, such as *machine* or *eunite2001*, our solution outperforms KNN and SVR in all the simulated AO scenarios. In general, its predominance is particularly notable when the swap rate, r, is intermediate. Although KNN is competitive in AO simulations with no (or few) swaps, its performance rapidly degrades as r increases. Its behavior is noticeably stable across domains and bag sizes. The increase in terms of RMSE is simply explained by the larger uncertainty associated to larger bags. Finally, SVR performs similarly to our solution based on linear models (LM). Its usual robust behavior against swaps is noteworthy: the larger the swap rate, r, the better its performance. A quick degradation of the performance of the linear models when r reaches 0.4 is discernible in several datasets (*mpg*, *boston_housing*, *stock*, *cpusmall*). In other domains such as *bank*, *bodyfat* or *mg*, as bag size m_i grows, the performance of SVR drops when r is small and that of LM also drops when r is large. Regarding the bag size, m_i, its comparative influence in these experiments is limited, as it affects all the methods similarly.

Figure 5 summarizes the results over domains. Regarding the swap rate, r, KNN only outperforms the rest of the methods in experiments without swaps, SVR stands out when the value of r is large, and the proposed LM solution is the best approach in scenarios with intermediate values of r. SVM is remarkably robust to swaps and its performance even improves as r increases. Our solution is robust for small swap rates ($r \leq 0.4$) but, as r grows, its performance draws a notable slope. The effect of the bag size, m_i, is consistent for all three methodologies: as larger values of m_i are considered, the aforementioned differences become more noticeable.

Once our solution and the state of the art approaches have been validated in simulated data, a real-world application of the AO framework is presented and these techniques are tested in real data.

4 Case study: marine litter beaching

Our interest in the AO framework is motivated by the study of a real problem where the response variable naturally shows aggregated outputs: the marine litter beaching prediction.

The objective is to build a model for predicting the waste accumulated during a single day on the beaches of seven different municipalities (1-3 beaches per locality) on the Basque coastline. Given such a prediction, local authorities could arrange the daily cleaning activities, set up resource (both workforce and machinery) allocation, or even take precautionary measures. Thus, each problem example represents a specific day, and it is described by a set of river-flow, metocean and environmental explanatory variables (see Table 2 for a com-

Table 2: Identifier and explanatory variables of the dataset.

Variable	Range	Description
Site	$\{1, \dots, 7\}$	Municipality-Beach(es) id.
Date	{01/01/2009, . . . , 12/31/2016}	Date of the measurements
Day	{Su, Mo, Tu, We, Th, Fr, Sa}	Day of the week
Season	{Spr,Sum,Aut,Win}	Season of the year
InputA-D	$[0, 13.7]$ (m)	Max. significant wave height, sections 1-4
InputE	$[0.80, 448.4]$ (m^3/s)	Water flow of Bidasoa river
InputF	$[1.19, 227.5]$ (m^3/s)	Water flow of Urumea river
InputG	$[2.71, 715.9]$ (m^3/s)	Water flow of Oria river
InputH	$[0.94, 286.4]$ (m^3/s)	Water flow of Urola river
InputI	$[0.58, 242.6]$ (m^3/s)	Water flow of Deba river
InputJ	$[0.56, 20.3]$ (m^3/s)	Water flow of Artibai
InputK	$[0.57, 9.5]$ (m)	Max. significant wave at Zarautz
InputL	$[3.45, 8.7]$ (m)	Max. coastal flooding level at Zarautz
InputM-P	$[0, 1381.0]$ (m)	Drag, sections 1-4
InputQ-T	$[0, 1734.9]$ (m)	Max. drag at Zarautz, sections 1-4

plete description). During the whole period 2009-2016, daily observations of the explanatory variables were automatically collected by means of strategically located sensors: the Bilbao-Vizcaya buoy[2], gauging stations along Gipuzkoan rivers[3] and the Zarautz weather station[4]. Due to the importance of wind-wave direction and beach orientation in litter beaching [33, 2, 10], the explanatory variables *wave height* and *drag* have been divided into four sections: (1) from Northwest to Northeast, (2) from Northeast to Southeast, (3) from Southeast to Southwest, and (4) from Southwest to Northwest. Regarding the response variable in this domain, historical data about waste accumulation is available for the same period 2009-2016. However, the measurement of waste accumulation involves two issues.

Firstly, the output is usually aggregated for consecutive days. The real amount of waste accumulated in each locality can only be measured as it is removed by the beach cleaning service, and cleaning all the beaches everyday is economically unfeasible. Thus, there is no record of waste accumulation for the days on which the service did not visit the beach. Note that, due to this factor, there are two types of examples depending on the frequency of the visits of the cleaning service to a municipality. On the one hand, if a beach is cleaned on (at least) two consecutive days, since waste has undoubtedly accumulated during the previous 24 hours, examples (days) with certain measurements of waste accumulation (a.k.a., fully supervised examples) are available. On the other hand, if a beach is not cleaned on consecutive days, the waste accumulates for several days until the cleaning service collects and measures it (bag of examples with AO). Thus, as certain waste measurements are occasionally available, one may question to which extent the examples with aggregated outputs enhance the performance of a model just learnt with these fully supervised examples.

[2] `http://www.puertos.es/en-us/oceanografia/Pages/portus.aspx`

[3] `http://www.gipuzkoahidraulikoak.eus/es/datos-tiempo-real`
`http://www.chcantabrico.es`

[4] `http://www.euskalmet.euskadi.eus/s07-5853x/es/meteorologia/lectur.apl`

The second issue involves the existence of natural processes, such as the tide, wind or waves [39, 2], that periodically clean beaches. When these natural dynamics are taken into account, it can be assumed that the aggregated output measured after a period without cleaning services would be lower than the total amount that would have been measured if cleaning had been carried out on a daily basis. Under this hypothesis, disaggregating the AO among all the examples of the bag (sequence of days from the last waste removal-measurement) is not accurate. To use the AO information fairly, it would be necessary to estimate the examples which really contribute to the AO of each bag (the days in which the removed waste was really deposited).

To sum up, in this case study we aim to show the contribution of an AO approximation to the problem of waste accumulation estimation, taking into account two important factors: (i) the possible existence of natural cleaning processes, and (ii) the possible presence of enough certain measurements that would make an AO approach redundant. To this end, we have carried out a complete set of experiments specifically designed to answer these questions.

4.1 Experiments

Aiming to answer both aforementioned objectives, different experimental scenarios are generated. For each municipality separately, different experimental scenarios work on different subsets of data for model training and evaluation.

On the one hand, the data is separated by year with the objective of evaluating the effect of the amount of training data. Remember that the original dataset comprises, for each municipality, daily samples for 8 consecutive years (i.e., 2920 examples per locality). Thus, for each municipality, 8 different datasets comprising all the examples gathered before or during (i.e., not after) each specific year have been generated. That is, the first subset ($\leq$2009) includes the examples of year 2009 and is, in fact, a subset of all the subsequent divisions. The second subset ($\leq$2010) comprises the examples of 2009-2010. The last subset ($\leq$2016) includes all the 2920 examples.

On the other hand, in order to test the idea of the existence of natural cleaning dynamics, large bags are shortened to a maximum bag size, $\breve{m}_i$. Specifically, 7 different maximum bag sizes, $\breve{m}_i = \{1, \ldots, 7\}$, have been tested. In bags larger than $\breve{m}_i$, the excess examples are removed (in practice, the last $\breve{m}_i$ examples of each bag remain since examples are chronologically ordered). The idea is that the waste present on the beach long before collection would have already been removed by the alleged natural dynamics. An enhanced performance of an AO solution using a specific $\breve{m}_i$ (among other $\breve{m}_i$ values with limited performance) might indicate the average amount of time that those natural dynamics need to remove litter.

All 7 maximum bag sizes where applied to all the 8 year subsets (examples up to and including that year, $\leq year$) to set 56 different experimental scenarios. All of them were generated and tested for all the 7 beaches. However, for the sake of simplicity, in this paper only results for municipalities #3 (Zumaia) and #7 (Donostia) are analyzed. The results for the rest of localities are collected in

Table 3: Training examples in Zumaia (Gipuzkoa) for different scenarios. Division per column shows subsets of examples gathered before or during each year (not after). Each row shows the number of examples in the bags for a different maximum bag size, $\breve{m}_i$. The last row shows the number of fully supervised examples, which is not affected by $\breve{m}_i$.

	≤year							
$\breve{m}_i$	2009	2010	2011	2012	2013	2014	2015	2016
1	70	135	205	268	336	390	401	468
2	140	270	410	536	672	780	802	936
3	195	369	563	746	926	1081	1114	1290
4	230	439	670	899	1113	1301	1345	1552
5	259	498	761	1031	1273	1483	1538	1770
6	277	542	832	1134	1399	1627	1693	1944
7	291	572	876	1200	1477	1720	1796	2063
fully sup.	41	108	145	154	183	205	206	229

Table 4: Training examples in Donostia (Gipuzkoa) for different scenarios. Division per column shows subsets of examples gathered before or during each year (not after). Each row shows the number of examples in the bags for a different maximum bag size, $\breve{m}_i$. The last row shows the number of fully supervised examples, which is not affected by $\breve{m}_i$.

	≤year							
$\breve{m}_i$	2009	2010	2011	2012	2013	2014	2015	2016
1	73	133	181	241	310	374	418	480
2	146	266	362	482	620	748	836	960
3	184	340	463	613	791	951	1069	1232
4	204	381	522	694	893	1076	1218	1407
5	213	404	560	747	959	1160	1320	1529
6	217	417	584	781	1001	1215	1390	1609
7	219	424	600	807	1031	1254	1440	1665
fully sup.	144	297	455	580	698	813	952	1082

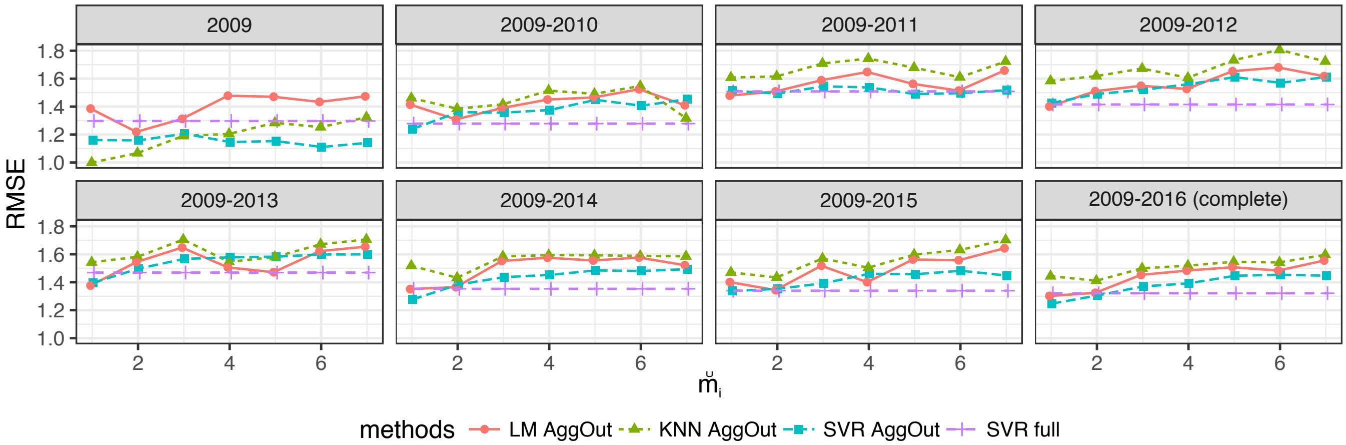


Figure 6: Results in terms of root mean square error of linear, KNN and SVR models learnt from the AO data of the beaches of Zumaia (Gipuzkoa). SVR models learnt only with the available supervised examples are used as a baseline. Each subfigure shows the results of experiments on a different subset (examples collected ≤ year) pruning the large bags by using an increasing maximum value for bag size, $\breve{m}_i = \{1, 2, \ldots, 7\}$.

the supplementary material document on the website associated with this study, together with all the numerical results of this set of experiments. Tables 3 and 4 show, for Zumaia and Donostia respectively, the size of the different training datasets used in each experimental scenario. In each row, the dataset sizes obtained are the product of applying a specific maximum bag size, $\breve{m}_i$, to the different year subsets ($\leq year$). In the last row, the amount of fully supervised examples (note that $\breve{m}_i$ does not affect this number) is shown.

As explained in the previous section, model validation is not feasible by means of standard techniques when only aggregated outputs are available. However, in this domain, we do have access to fully supervised examples (last row in Tables 3 and 4): the examples describing consecutive days when waste was removed daily. A 10 × 5-fold cross validation was used for model evaluation in all these experiments. The folds were built only with the fully supervised examples, and the examples in the bags (with aggregated outputs) were always considered for model training. For all the experiments, three regressors were learnt: a linear model with our AO proposal, and Musicant et al.'s KNN and SVR. Additionally, a SVR is also learnt using only the fully supervised examples (without the AO approximation) to set a baseline for performance comparison.

The results of the experiments with data from the municipality of Zumaia are displayed in Figure 6. In this case, the model learnt only with fully supervised examples (without AO) poses a competitive baseline. With only the inclusion of the 2010 data (second subfigure of Figure 6), the fully supervised

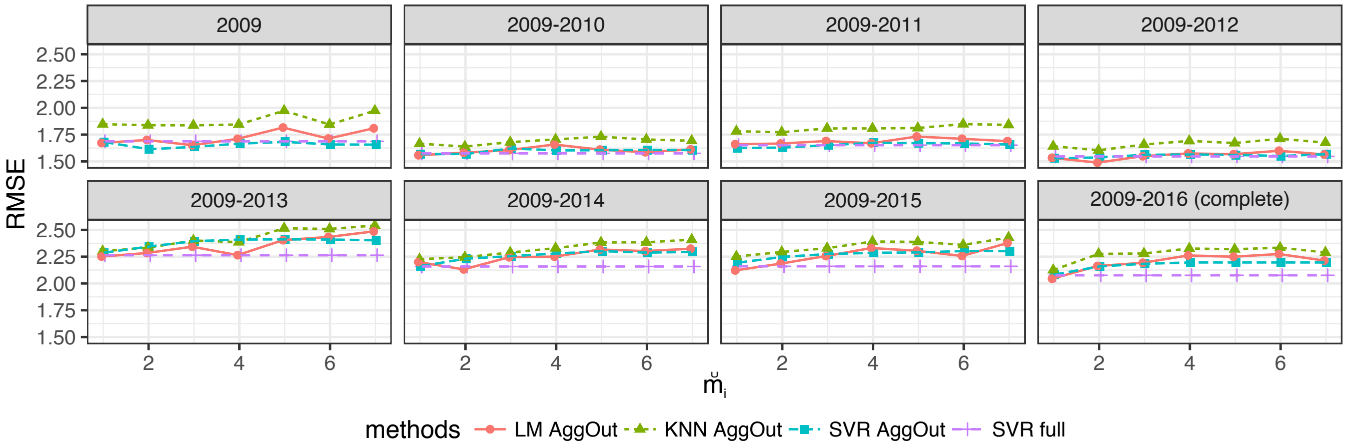


Figure 7: Results in terms of root mean square error of linear, KNN and SVR models learnt from the AO data of the beaches of Donostia (Gipuzkoa). SVR models learnt only with the available supervised examples are used as a baseline. Each subfigure shows the results of experiments on a different subset (examples collected ≤ year) pruning the large bags by using an increasing maximum value for bag size, $\breve{m}_i = \{1, 2, \ldots, 7\}$.

approach is already the best performance method in almost every experimental scenario. As more examples are used for training and validation (other years are incorporated), the difference with respect to the AO methodologies is generally clearer. The performance of all four approaches is affected when the 2011 data is included. After a manual examination of the 2011 data, this can be considered an atypical year with extremely unusual high values of collected litter (as well as high values of other variables). Apart from this, the behavior is steady: the larger the training dataset, the better the results. In the first years, when data is still scarce, the real contribution of the AO approximation is noticeable. For example, the SVR model with the AO approach outperforms the fully supervised SVR in all the scenarios of the 2009 subset. This shows the superior performance of AO methods when fully supervised examples are scarce (see Table 3). LM and, mainly, KNN also overcome the baseline in different scenarios. Globally, the best performing AO approach is SVR, and KNN the worst one.

The influence of the maximum bag size, $\breve{m}_i$, is also noteworthy. The performance of SVR models learnt with AO improves as $\breve{m}_i$ decreases (mainly, from subsets containing data from 2012 on). A similar trend can also be observed with LM and KNN models. However, this behavior is less clear as the maximum bag size drops below 5. It is worth noting that in almost all the subfigures of Figure 6, the performance of LM and KNN has a local minimum in the experimental setting with a maximum bag size of 2 and/or 4-5. In many cases, the performance in these points is not worse than when a maximum bag size $\breve{m}_i = 1$ (i.e., no bags are used) is imposed. These results are also competitive regarding the fully supervised baseline approach. These valleys could identify scenarios where the decision of shortening the bags is particularly appropriate.

The experimental results with data from Donostia show similar trends (see Figure 7). As previously mentioned, the error of the four approaches monotonically decreases as the training dataset is enlarged (inclusion of data of additional years in each subfigure). Only the inclusion of the data of year 2013 produces the opposite effect. Like 2011 data from Zumaia, this was an atypical year on the beaches of Donostia. In this case, the difference between methods is slight and when AO approaches outperform the baseline, the gain is limited. This is probably an effect of the larger number of fully supervised training examples available for this beach (see Table 4). From the subsets including the 2013 year data and on, the SVR models learnt only with fully supervised data perform better than the AO approaches. However, in the first four subsets, AO approaches (mainly LM and SVR) are competitive or overcome the results of the fully supervised baseline. KNN is again the worst regressor and the differences between SVR and LM are negligible. In the case of this beach, the performance of the AO approaches shows the referred performance *valleys* too. As in the previous case, experiments with maximum bag size of 2 or 4 show enhanced performance.

To summarize, the experimental results of these two beaches (and those in the supplementary material) show that, when available, fully supervised training data masks the contribution of the aggregated outputs. When a sufficiently large amount of supervised examples is available, AO might not be necessary. In the experiments with the smallest subsets of (supervised) data, the contribution of the AO approach is noticeable. The existence of points with maximum bag size larger than 1 (usually, 2 or 4) where the performance of AO approaches reach or overcome that of the fully supervised approach could imply that the accumulated waste does not last longer than 4-5 days without being removed by the alleged natural dynamics such as the tide.

5 Discussion

Throughout this paper, we have shown that learning regression models with aggregated outputs is feasible. One may benefit from using a specific learning technique that exploits the information of the AO. Our methodological proposal and the state-of-the-art techniques of Musicant et al. [27] provide a set of tools to robustly deal with the whole spectrum of AO scenarios. The experimental results show the usefulness of these approaches with both synthetic and real data of a case study in marine litter beaching. However, if a sufficiently large number of fully supervised examples (bags with $m_i = 1$) is available, the AO approach might be unnecessary. However, the specific number of supervised examples required depends on the domain. When the number of fully supervised examples is not large, a model learnt from AO has shown to be competitive with a standard solution. This fact is particularly revealing and motivates the use of specific AO techniques in scenarios where fully supervised examples are scarce. It is interesting to know that other beaches in the Gipuzkoan Bay of Biscay, under the charge of the same cleaning service, were not included in this study as model validation was not possible due to the lack of fully supervised cases. The proposed approach is especially suitable for those infrequently visited beaches.

Remember that in these experiments the set up of the techniques was not modified. Our proposal always initializes the disaggregated response vector $\boldsymbol{y}'$ by Equation 5. The hyper-parameters of both KNN and SVR were set up as suggested by Musicant et al. [27]. In all the cases, the initialization and model parameters could be tuned to optimize the results.

Regarding the case study on marine litter beaching, it is important to note that, in the different experiments, a single maximum bag size, $\breve{m}_i$, was imposed for all the bags. Thus, two specific points ($\breve{m}_i = \{2, 4\}$) have been identified where the learnt models show enhanced performance, providing an idea of the behavior (frequency) of those dynamics that naturally clean the beaches. However, nature rarely shows such constant behavior. The temporal point represented by this experimental parameter is likely different for each beach/municipality and, even, each bag (group of days). Moreover, our approximation to this case study does not analyze the possible temporal correlations among explanatory and response variables. It is reasonable to think that the environmental conditions of a specific day may have an effect on the state of the beaches a few days later. An approximation based on time series seems a reasonable step forward.

6 Conclusions and future work

In this paper, an iterative approach to learn regression models from aggregated outputs has been proposed. In this framework, the response value is only provided in an aggregated way for subgroups of examples. The proposed methodology is implemented to learn linear models which, in spite of their popularity, have never been adapted to the AO framework. However, it is general enough to be straightforwardly extended to learn other types of (non-)linear models. The benefits of exploiting the supervision available in the aggregated outputs has been shown in comparison with two different baselines. In spite of its simplicity, our proposal is competitive with respect to a baseline and previously presented methods.

In the second part of the paper, the problem of marine litter beaching prediction, which naturally fits the AO framework, has been approached for the first time by means of techniques of this weakly supervised paradigm. A case study on real data from the Basque coast has been analyzed. Although this approximation is not necessary when enough supervised data is available, the models learnt from AO are competitive and particularly useful when fully supervised data is scarce. Finally, an interesting insight could apparently be drawn from the results of this case study: the waste accumulated on a beach was probably deposited during the previous two to four days. In the future, it would be of interest to study the effect of natural cleaning dynamics and to measure the time that the accumulated waste lasts on a beach if it is not removed. With the results of such a study, our approximation could be properly tuned and exploited.

Other future works would include the redesign of our proposal to learn other types of regression models by adapting Equations 6 and 7. A completely different approach to marine litter beaching prediction could be to consider a multivariate time series approach where the objective is to complete an occasionally observed variable, the amount of accumulated waste.

Acknowledgments

J.H.-G., I.I. and J.A.L. are partially supported by the Basque Government (IT609-13, Elkartek BID3A), the Spanish Ministry of Economy and Competitiveness (TIN2016-78365-R). J.A.L. is also supported by BERC program 2018-2021 (Basque Government) and Severo Ochoa Program SEV-2017-0718 (Spanish Ministry of Economy and Competitiveness). I.G., O.C.B. and J.A.F. are supported by the European Union (LIFE LEMA project, LIFE15/ENV/ES/000252). I.G. has also benefited from a grant from the Training of Technologists Program of the Department of Economic Development and Infrastructures of the Basque Government. J.A.F. work is funded by the Gipuzkoa Talent Fellowships by Gipuzkoa Provincial Council. This paper is the contribution no. XXX AZTI-Marine Research. The authors would like to thank the Department of the Environment and Water Management Works of Provincial Council of Gipuzkoa for providing the waste data on the beaches from Gipuzkoa, and the anonymous reviewers for their helpful comments.

References

[1] J. Bellas, J. Martínez-Armental, A. Martínez-Cámara, V. Besada, and C. Martínez-Gómez. Ingestion of microplastics by demersal fish from the spanish atlantic and mediterranean coasts. *Marine Pollution Bulletin*, 109(1):55–60, 2016.

[2] L. C. Blickley, J. J. Currie, and G. D. Kaufman. Trends and drivers of debris accumulation on maui shorelines: Implications for local mitigation strategies. *Marine Pollution Bulletin*, 105(1):292–298, 2016.

[3] R. Brouwer, D. Hadzhiyska, C. Ioakeimidis, and H. Ouderdorp. The social costs of marine litter along european coasts. *Ocean & Coastal Management*, 138:38–49, 2017.

[4] C.-C. Chang and C.-J. Lin. LIBSVM: a library for support vector machines, 2011.

[5] S. Chen, B. Liu, M. Qian, and C. Zhang. Kernel k-means based framework for aggregate outputs classification. In *IEEE International Conference on Data Mining Workshops (ICDM Workshops 2009)*, pages 356–361, 2009.

[6] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[7] J. G. Derraik. The pollution of the marine environment by plastic debris: a review. *Marine pollution bulletin*, 44(9):842–852, 2002.

[8] L. M. Dery, B. Nachman, F. Rubbo, and A. Schwartzman. Weakly supervised classification in high energy physics. *Journal of High Energy Physics*, 2017(5):145, 2017.

[9] S. Deudero and C. Alomar. Mediterranean marine biodiversity under threat: Reviewing influence of marine litter on species. *Marine Pollution Bulletin*, 98(1):58–68, 2015.

[10] C. Eriksson, H. Burton, S. Fitch, M. Schulz, and J. van den Hoff. Daily accumulation rates of marine debris on sub-antarctic island beaches. *Marine Pollution Bulletin*, 66(1):199–208, 2013.

[11] B. Fish and L. Reyzin. On the complexity of learning from label proportions. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 19–25, 2017.

[12] E. M. Foundation and M. . Company. The new plastics economy: Rethinking the future of plastics. http://www.ellenmacarthurfoundation.org/publications/the-new-plasticseconomy-rethinking-the-future-of-plastics, 2016.

[13] A. Frank and A. Asuncion. UCI machine learning repository, 2010.

[14] F. Galgani, D. Fleet, J. Van Franeker, S. Katsanevakis, T. Maes, J. Mouat, L. Oosterbaan, I. Poitou, G. Hanke, R. Thompson, E. Amato, A. Birkun, and C. Janssen. *Marine Strategy Framework directive - Task Group 10 Report. Marine litter.* Office for Official Publications of the European Communities, 2010. EUR 24340 EN - 2010.

[15] S. Gall and R. Thompson. The impact of debris on marine life. *Marine Pollution Bulletin*, 92(1):170–179, 2015.

[16] K. Hall. *Impacts of Marine Debris and Oil: Economic and Social Costs to Coastal Communities.* Kommunenes Internasjonale Miljøorganisasjon, Shetland: KIMO, 2000.

[17] J. Hernández-González, I. Inza, L. Crisol-Ortiz, M. A. Guembe, M. J. Iñarra, and J. A. Lozano. Fitting the data from embryo implantation prediction: Learning from label proportions. *Statistical Methods in Medical Research*, 27(4):1056–1066, 2018.

[18] J. Hernández-González, I. Inza, and J. A. Lozano. Learning Bayesian network classifiers from label proportions. *Pattern Recognition*, 46(12):3425–3440, 2013.

[19] J. Hernández-González, I. Inza, and J. A. Lozano. Weak supervision and other non-standard classification problems: A taxonomy. *Pattern Recognition Letters*, 69:49–55, 2016.

[20] D. Hübner, T. Verhoeven, K. Schmid, K.-R. Müller, M. Tangermann, and P.-J. Kindermans. Learning from label proportions in brain-computer interfaces: Online unsupervised learning with guarantees. *PloS one*, 12(4):e0175856, 2017.

[21] H. Kück and N. de Freitas. Learning about individuals from group statistics. In *Proceedings of the 21st Conference in Uncertainty in Artificial Intelligence (UAI)*, pages 332–339, 2005.

[22] D. Laist. Marine debris entanglement and ghost fishing: A cryptic and significant type of bycatch? In *Proceedings of the Solving Bycatch Workshop*, pages 33–39, Seattle, WA, USA, 1995.

[23] D. W. Laist. *Impacts of Marine Debris: Entanglement of Marine Life in Marine Debris Including a Comprehensive List of Species with Entanglement and Ingestion Records*, chapter 8, pages 99–139. Springer New York, New York, NY, USA, 1997.

[24] L. López-López, I. Preciado, J. M. González-Irusta, N. L. Arroyo, I. Muñoz, A. Punzón, and A. Serrano. Incidental ingestion of meso- and macro-plastic debris by benthic and demersal fish. *Food Webs*, 14:1–4, 2018.

[25] J. M. Adey, I. Smith, R. Atkinson, I. Tuck, and A. Taylor. ‘ghost fishing’ of target and non-target species by norway lobster nephrops norvegicus creels. *Marine Ecology Progress Series*, 366:119–127, 2008.

[26] A. McIlgorm, H. F. Campbell, and M. J. Rule. The economic cost and control of marine debris damage in the asia-pacific region. *Ocean & Coastal Management*, 54(9):643–651, 2011.

[27] D. R. Musicant, J. M. Christensen, and J. F. Olson. Supervised learning by training on aggregate outputs. In *Proceedings of the 7th IEEE International Conference on Data Mining (ICDM)*, pages 252–261, 2007.

[28] S. Newman, E. Watkins, A. Farmer, P. Brink, and J.-P. Schweitzer. The economics of marine litter. In M. Bergmann, L. Gutow, and M. Klages, editors, *Marine anthropogenic litter*, pages 367–394. Springer, 2015.

[29] K. R. Nicastro, R. Lo Savio, C. D. McQuaid, P. Madeira, U. Valbusa, F. Azevedo, M. Casero, C. Loureno, and G. I. Zardi. Plastic ingestion in aquatic-associated bird species in southern portugal. *Marine Pollution Bulletin*, 126:413–418, 2018.

[30] G. Patrini, R. Nock, P. Rivera, and T. Caetano. (almost) no label no cry. In *Proceedings of Advances in Neural Information Processing Systems (NIPS 27)*, pages 190–198, 2014.

[31] M. Pérez-Ortiz, P. A. Gutiérrez, M. Carbonero-Ruz, and C. Hervás-Martnez. Adapting linear discriminant analysis to the paradigm of learning from label proportions. In *Proceedings of the 2016 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1–7, 2016.

[32] M. E. Portman and R. E. Brennan. Marine litter from beach-based sources: Case study of an eastern mediterranean coastal town. *Waste Management*, 69:535–544, 2017.

[33] M. Prevenios, C. Zeri, C. Tsangaris, S. Liubartseva, E. Fakiris, and G. Papatheodorou. Beach litter dynamics on mediterranean coasts: Distinguishing sources and pathways. *Marine Pollution Bulletin*, 2017. in press.

[34] Z. Qi, F. Meng, Y. Tian, L. Niu, Y. Shi, and P. Zhang. Adaboost-llp: A boosting method for learning with label proportions. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–12, 2017. early access.

[35] Z. Qi, B. Wang, F. Meng, and L. Niu. Learning with label proportions via npsvm. *IEEE transactions on Cybernetics*, 47(10):3293–3305, 2017.

[36] N. Quadrianto, A. J. Smola, T. S. Caetano, and Q. V. Le. Estimating labels from label proportions. *Journal of Machine Learning Research*, 10:2349–2374, 2009.

[37] S. Rech, V. Macaya-Caquilán, J. Pantoja, M. Rivadeneira, D. J. Madariaga, and M. Thiel. Rivers as a source of marine litter – a study from the se pacific. *Marine Pollution Bulletin*, 82(1):66 – 75, 2014.

[38] S. Rüping. SVM classifier estimation from group probabilities. In *Proceedings of the 27th International Conference on Machine Learning (ICML 2010)*, pages 911–918, 2010.

[39] P. G. Ryan, A. Lamprecht, D. Swanepoel, and C. L. Moloney. The effect of fine-scale sampling frequency on estimates of beach litter accumulation. *Marine Pollution Bulletin*, 88(1):249–254, 2014.

[40] C. Schmidt, T. Krauth, and S. Wagner. Export of plastic debris by rivers into the sea. *Environmental Science & Technology*, 51(21):12246–12253, 2017.

[41] O. Setälä, J. Norkko, and M. Lehtiniemi. Feeding type affects microplastic ingestion in a coastal invertebrate community. *Marine Pollution Bulletin*, 102(1):95–101, 2016.

[42] STAP. *Marine Debris as a Global Environmental Problem: Introducing a solutions based framework focused on plastic.* A STAP Information Document. Global Environment Facility, Washington, DC, USA, 2011.

[43] M. Stolpe and K. Morik. Learning from label proportions by optimizing cluster model selection. In *Proceedings of the European conference on Machine learning and knowledge discovery in databases (ECML/PKDD 2011)*, volume 3, pages 349–364, 2011.

[44] T. Sun, D. Sheldon, and B. O'Connor. A probabilistic approach for learning with label proportions applied to the US presidential election. In *2017 IEEE International Conference on Data Mining (ICDM)*, pages 445–454, 2017.

[45] L. Torgo. The LIACC repository, 2000.

[46] A. Unger and N. Harrison. Fisheries as a source of marine debris on beaches in the united kingdom. *Marine Pollution Bulletin*, 107(1):52–58, 2016.

[47] F. X. Yu, D. Liu, S. Kumar, T. Jebara, and S. Chang. $\propto$SVM for learning with label proportions. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, pages 504–512, 2013.