Collaborative Wind Power Forecast

Vânia Almeida[1] and João Gama[1,2]

[1] LIAAD - INESC TEC, University of Porto, Porto, Portugal
vania.g.almeida@inescporto.pt,
[2] Faculty of Economics, University of Porto, Portugal
jgama@fep.up.pt

Abstract. There are several new emerging environments, generating data spatially spread and interrelated. These applications reinforce the importance of the development of analytical systems capable to sense the environment and receive data from different locations. In this study we explore collaborative methodologies in a real-world problem: wind power prediction. Wind power is considered one of the most rapidly growing sources of electricity generation all over the world. The problem consists of monitoring a network of wind farms that collaborate by sharing information in a very short-term forecasting problem. We use an auto-regressive integrated moving average (ARIMA) model. The Symbolic Aggregate Approximation (SAX) is used in the selection of the set of neighbours. We propose two collaborative methods. The first one, based on a centralized management, exchange data-points between nodes. In the second approach, correlated wind farms share their own ARIMA models. In the experimental work we use 1 year data from 16 wind farms. The goal is to predict the energy produced at each farm every hour in the next 6 hours. We compare the proposed methods against ARIMA models trained with data of each one of the farms and with the persistence model at each farm. We observe a small but consistent reduction of the root mean square error (RMSE) of the predictions.

Keywords: Wind Power, Time Series Analysis, Collaborative Forecast, Correlation, Arima

1 Introduction

Emerging environments generate data spatially spread and interrelated. These applications reinforce the importance of the development of analysis systems capable to sense the environment and receive data from different locations [1]. The capability to integrate the overall set of information available can be meaningful and can be used in the development of proper adaptive data analysis algorithms.

Wind power is considered one of the most rapidly growing sources of electricity generation all over the world [2]. The main problems remain on the modelling of the wind turbine output [3] and on the development of accurate wind power forecast methodologies, capable to deal with the uncertain and variability of this resource. The suitability of a forecasting model is determined by the forecasting horizon which is the time ahead for which the forecast is made [4], being mainly separated into very short-term (30min-6hrs), short term (up to 72hrs ahead) and long term forecasting (several days ahead) [2, 5]. Statistical methods are commonly used for short-term wind forecast, taking as input the past values from the forecast variable. The most popular models are auto-regressive moving average (ARMA) models and their variants, *e.g.*, Auto-Regressive Integrated Moving Average (ARIMA), seasonal- and fractional-ARIMA and ARIMA with exogenous input (ARMAX or ARX). The development of prediction tools is not a new subject, and there is a considerable number of important contributions on this topic [6, 7].

Motif discovery commonly used to reveal trends, relationships, and anomalies can provide some guidance on the analysis of correlations between wind farms. This subject was studied by Kamath and Fan (2012) [8] using the Symbolic Aggregation Algorithm (SAX) [9]. In this work, it was discussed the role of motifs in scheduling operations.

The evolution of weather fronts over an extended area generates dependencies between power generations at different locations that can be useful to improve forecast methodologies. It was demonstrated that the combination of Numerical Weather Predictions (NWP) from different stations leads to the error decrease [10]. Berdugo et al. [11] described a collaborative short term forecasting methodology for photovoltaic problem. The results indicate the improvement of the forecast error when collaboration among sites is employed, comparatively to standard reference methods. A similar methodology for short-term wind speed prediction using both temporal and spatial characteristics also demonstrated the relevance of the spatio-temporal prediction tasks [12]. The forecasting task for geo-referenced time series also demonstrates the effectiveness of spatial and temporal ARIMA modelling with respect to univariate time series [13].

Although ARIMA is broadly used in time series analysis, there are few few studies considering the spatially correlation among data from different locations. This paper proposes a collaborative approach where wind-farms share data. We start by identifying correlations, trends, and patterns between farms, and exploit these correlations for optimizing predictions. The main contribution is the development of a collaborative wind power forecast approach, considering the interrelation among neighbour farms. The preliminary selection of potentially correlated farms consisted on the search for motifs using the SAX.

The organization of the paper is as follows. In Section 2 the collaborative forecast methodology is described. Experimental validation on real wind power dataset is presented in Section 3. The final section concludes the paper, including foreseen future work.

2 Collaborative Forecast for Network Data

A collaborative prediction approach applied to wind power forecast is proposed. However, this approach is no dependent on this particular application and can be seen as a general approach to other real world domains that have similar forecasting problems with the same type of network data. The application to sensor network problems lead us to consider the computational power a problem, even being aware that for this specific application be a less important requirement.

The goal consists of monitor a network of N synchronized sites (wind farms), numbered *i=1,2,...,N*. Each site has a set (NG_i) of correlated neighbour sites that collaborate to optimally fit the wind power forecast, at a 6-hours ahead horizon. The expected output is to minimize the forecast error of a site i, sharing relevant information but using minimum communication costs.

2.1 Finding Motifs

The preliminary selection of the potentially correlated neighbours to include in collaborative wind power model was performed searching for recurring motifs in historical data. A subsequence that repeats at least once is a motif. For the evaluation of the relationship between two subsequences, a distance measure must be used, as well as a match threshold. It is important to consider that a re-occurrence of the subsequence needs not to be exact for it to be considered as a motif. To map into a lower dimensional space, the SAX algorithm proposed by Lin *et al.* [14] was adopted.

The relation of patterns for different wind farms with different installed capabilities is a difficult task. So, before to apply SAX, data was scaled to maximum installed capacity, assuring the minimization of the distance between subsequences. This task is essential for the definition of the similarity threshold value.

2.2 Computation from Correlation Matrices

The computation of spatial and temporal correlation plays an important role in distributed environments [15], being possible to determine the strength of the influence of the distributed data. Along this work, different types of networks (and thus correlation measures) describing interactions between nodes are considered. The Pearson correlation is used in centralized management, while distributed approaches use the dot-product analysis.

Pearson correlation measures the linear correlation (dependence) between two variables x and y, giving a value between $+1$ and -1 inclusive, where 1 is total positive correlation, 0 is no correlation, and -1 is total negative correlation.

The dot product is also considered as a correlation metric, allowing to measure how closely two feature vectors are related. It is defined as the cosine of the angle of a paired data represented as vectors, $x.y = |x|\ |y|\ cos(\theta)$. For each single site, we compute the inner product between consecutive subsequences of fixed length (6-hours in this case). Both methods require the determination of a minimum threshold for the correlation coefficient.

2.3 Persistence

The persistence is a common used baseline prediction model. It considers that the wind power in the next time step is the same as occurred in the present time.

A known generalization was used, considering the prevision at time instant t for a look-ahead time $t+k$ ($\hat{p}_{t+k|t}$) the average value of the last n observations ($n = 6$ hours in this case), being defined as follows:

$$\hat{p}_{t+k|t} = \frac{1}{n}\sum_{i=0}^{n-1} p_{t-i}$$

2.4 ARIMA Modelling

The ARIMA modelling approach was introduced by Box and Jenkins (1976) [16] to analyse stationary univariate time series, taking as input the past values from the forecast variable. Along this work all models were implementation in R using the *forecast* package.

Three models were implemented, a ARIMA reference model (RefARIMA) comprised the train for the historical observations of each one of the farms, using the *auto.arima* function, and two collaborative models. The collaborative models were denominated CentARIMA and DistARIMA. CentARIMA is a model based on centralized management that employs exchange of the values of time series between nodes. The another one, DistARIMA, takes into account the limited computational power associated to the sensor network topologies. In this case, the correlated wind farms share their own ARIMA models.

Centralized Approach The first idea to solve the forecast problem consisted on the combination of correlated subsequences from the network data. The Pearson correlation is used to search for correlated sequences, considering the NG set. A threshold *thd* is defined to considerate a correlation ($thd > 0.7$). Wind power production at a given site i is a weighted linear combination of past production values at a set of neighbour sites. The *auto.arima* model is performed for the weighted time series (w_i) at each site. From the analysis of the correlation value, it is clear that a high correlation value could arise from data at different amplitude scales. The prediction values need to be adjusted to the correct baseline level. The adjustment consists in the removal of the difference observed between the past 6-values (mean value) and the first prediction value. This algorithm is described in Alg. 1.

Distributed Approach For each wind farm, the past 2 subsequences of length k are used to compute the dot product. If the dot value is higher that the established threshold ($thd > 0.97$), the predicted values are computed normally, using the *auto.arima* function. Otherwise, being the dot product value lower than the acceptable, the correlated set of wind farms share their own ARIMA models. The final prediction is the weighted sum at each hour of the predicted values obtained for the N considered models. This methodology intends to avoid higher prediction errors that may occur when the actual situation is not correlated with the past, using information from the other farms that experienced similar conditions previously. This procedure is described in Alg. 2.

Algorithm 1: `CentARIMA`: Centralized ARIMA.

```
input  : $S_i$: Stream of wind power values for farm i
         $NG_i$: Set of correlated neighbour sites
         $k$: Length of sequences used in correlation
         $j$: Identification of past values
         $n$: Number of observations used to ARIMA train
         $thd$: Correlation threshold
output: 6-hours ahead wind power forecast
begin
    foreach $farm\ i$ do
        foreach $t \in S_i$ do
            $s_i \leftarrow$ Set of sequences ($< x_i(t-k-j), ..., x_i(t-j) >$) from $NG_i$
            Compute Pearson correlation A for the sequences in $s_i$
            if $correlation > thd$ then
                $A_{i,j} \leftarrow 1$
                $count_c \leftarrow count_c + 1$
            else
                $A_{i,j} \leftarrow 0$
            $w_i(\mathrm{t}) \leftarrow \frac{1}{count_c} \sum_{j \in s_i} x_i(t-j).A_{i,j}$
            if $t > n$ then
                Fit $auto.arima$ for $< w_i(t-n), ..., w_i(t) >$
                $\hat{x}_i(t+1), ..., \hat{x}_i(t+6) \leftarrow$ predicted data
                $\hat{x}'_i(t+1), ..., \hat{x}'_i(t+6) \leftarrow$ adjust predictions to amplitude scale
```

3 Experimental Setup

3.1 Data

For the experiments, we took data from 16 wind farms, distributed at different geographical sites. Data from one year of power production at a hourly-step are available. The set of neighbour farms was chosen based on the number of pairs and motifs occurrence at different lengths, using the SAX representation. The maximum time horizon was set up to 720 hrs (30 days).

3.2 Error Measure

The accuracy of the models is measured by the root mean squared error (RMSE), expressed as a percentage between $\hat{x}_t$ (the forecast at time t) and x_t (the observed value). The analysis was performed in a hour-ahead step until to 6-hours (eq. below).

$$RMSE = \sqrt{\frac{1}{6} \sum_{t=1}^{6} (\hat{x}_t - x_t)^2}$$

Algorithm 2: `DistARIMA`: Distributed ARIMA.

```
input  : $S_i$: Stream of wind power values for farm i
         $k$: Length of correlated sequences
         $n$: Number of observations used to ARIMA train
         $NG_i$: Set of correlated neighbour sites
         $N$: Length of NG set
         $thd$: Correlation threshold
output:  6-hours ahead wind power forecast
begin
    foreach $farm\ i$ do
        foreach $t \in S_i$ do
            Collect last 2 consecutive sensed data sequences of length k:
            $(x_1, ..., x_k$ and $y_1, ..., y_k)$
            Compute DOT=$< x_1, ..., x_k >.< y_1, ..., y_k >$
            if $DOT > thd$ then
                Run auto.arima function for the last n observations
                $(\hat{x}_{t+1}, ..., \hat{x}_{t+6})\leftarrow$ predicted data
            else
                Receive ARIMA model parameters from the $NG_i$ set
                Fit N ARIMA models for the last n observations
                $(\frac{1}{N}\sum_{i\in NG_i}\hat{x}_i(t+1), ..., \frac{1}{N}\sum_{i\in NG_i}\hat{x}_i(t+6))\leftarrow$ predicted data
```

3.3 Evaluation of the Predicted Methods

The evaluation was performed for the entire dataset (1 year) in a hourly-step, being the first n observations required to initialize the models.

Centralized Model The data analysis consisted on the training of an ARIMA model with 100 observations. The evaluation results are presented in Figure 1, using the RefARIMA with the same number of observations, as comparison. For all the farms, we observe lower prediction errors associated to the CentARIMA. The average decrease value is 0.56%. Using the Wilcoxon test, and considering a p-value< 0.01, the differences between models are significant for all the farms excluding the WF15 with $p = 0.37$.

We consider that the exchange of 100 observations for a large network is a number not acceptable in sensor networks. Several experiments for different data length were performed. Figure 2 shows that the historical data length is preponderant on the ARIMA model error. Large historical data length are associated to lower errors but implies more computation cycles and memory usage. The collaborative model, CentARIMA is more stable compared to the traditional univariate model RefARIMA. In this case, the number of historical observations has no prominent influence on the error value, up to less than 100 observations. On the other hand, the historical data length has a preponderant effect on the accuracy of RefARIMA that increases for the models using fewer historical data.

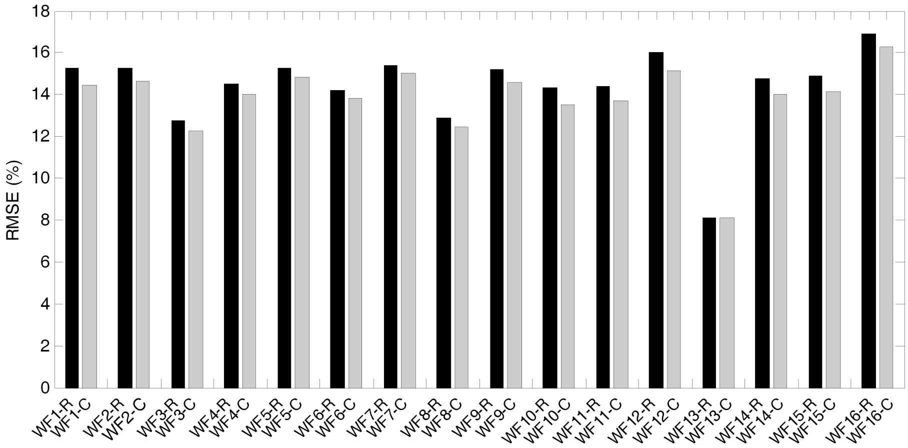

Fig. 1. RMSE values (8600 experiments) from RefARIMA (black bars) and CentARIMA (grey bars), trained with 100 observations for a horizon of 6-hours.

It is possible to conclude that the collaborative model presents competitive advantages, if the historical data length is a requirement, without compromising the error value and avoiding computation cycles and memory usage.

Distributed Model Some textbooks provide rules to minimum sample sizes for various time series models. In the case of ARIMA, 30 observations is often refereed as the minimum acceptable number. So, the DistARIMA model was implemented using 30 observations, being the results compared to the RefARIMA.

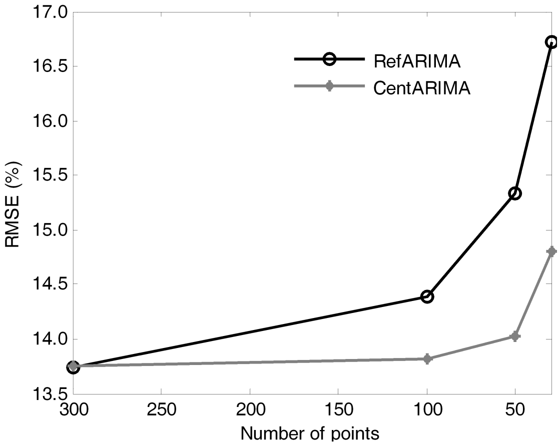

Fig. 2. RMSE error for different historical data length used on the ARIMA model train, at black the RefARIMA and at grey the CentARIMA simulations.

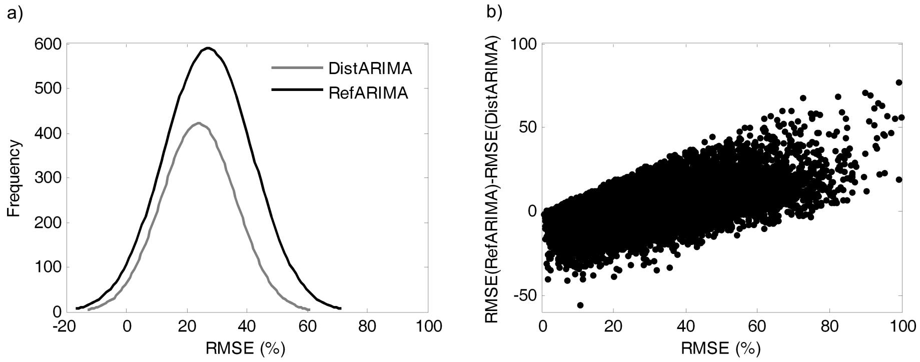

Fig. 3. a) RMSE values when *dot* < 0.97 for the RefARIMA (black) and DistARIMA (grey) models, being visible the lower error distribution for the DistARIMA. b) RMSE difference, being visible the improvement of the DistModel for higher RMSE values.

Results are presented in Figure 3, at the left panel is represented the RMSE error distribution for both models, and at the right the observed differences are plotted. It was observed an average decrease of 3.24% for the DistARIMA, considering the zones where dot product <0.97 (the predefined threshold). It is also visible at the right panel that the error associated to the DistARIMA decreases for zones where the absolute error is higher, such as expected. The Wilcoxon test was applied and results indicate significant differences between the models for all of the farms.

Comparison of the Models The comparison included persistence and ARIMA models trained with 100 points. Firstly, the RMSE is compared at each hour ahead. The performance of three of the farms is presented in Figure 4. The relevance of the collaborative approaches is exposed, with lower error values comparatively to the persistence that only outperforms (average for all farms)

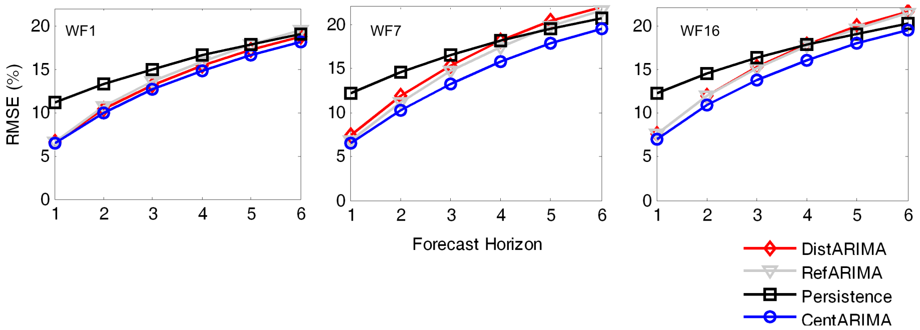

Fig. 4. Hour-ahead forecast for the persistence (black), RefARIMA (grey), CentARIMA (blue), DistARIMA (red) models for the WF1, WF7 and WF16.

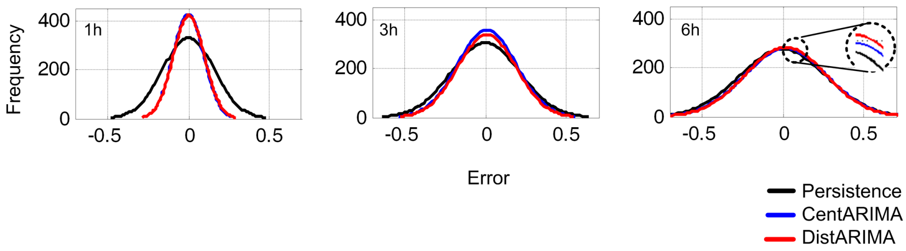

Fig. 5. Hour-ahead error measure for the persistence (black), CentARIMA (blue), DistARIMA (red) models for 1h, 3h and 6h.

one of the collaborative models (DistARIMA) for forecast horizons between $4-6$ hours. Comparing the ARIMA models with persistence, the improvement of DistARIMA is not so good comparatively to CentARIMA (average improvement of 0.38% *vs.* 2.46%, respectively).

We also present the analysis of the error distribution. Figure 5 points-out that no bias is present, considering all the models. For the persistence model in $1 - hour$ horizon is visible a wider dispersion comparatively to the ARIMA models. However, the difference is attenuated for $6 - hour$ horizon. Although, these numbers may seem relatively small, they have an interesting impact on the production costs.

4 Conclusions and Future Work

This paper discusses the advantages of a collaborative approach in short term wind power forecast. Two scenarios were tested, a centralized approach sharing time series between nodes and a distributed version that exchanges only the model parameters between nodes. It was observed RMSE decrease by 2.46% for the centralized and 0.38% for the distributed approach comparatively to the persistence values. These values result from 8600 experiments. In overall, a small but consistent RMSE reduction of the predictions was observed.

The work reported in this paper opens several directions of future research. The most obvious direction lies on the challenge of selecting the correlation threshold for that the forecast error is minimized. Further studies include the analysis of the influence of several parameters on the quality of results, such as k, NG_i, N, thd. Finally, research on other domains where data are network distributed is being planned.

Acknowledgements . This work was supported by projects Smartgrids (NORTE-07-0124-FEDER-000056) and Sibila (NORTE-07-0124-FEDER-000059) financed by ON.2−O Novo Norte, under the National Strategic Reference Framework, through the Development Fund, and by national funds, through FCT, and by European Commission through the project MAESTRA (Grant number ICT-2013-612944). The authors also acknowledge the support of Prewind.

References

1. M. May and L. Saitta, "Introduction: The challenge of ubiquitous knowledge discovery," in *Ubiquitous Knowledge Discovery* (M. May and L. Saitta, eds.), vol. 6202 of *Lecture Notes in Computer Science*, pp. 3–18, Springer Berlin Heidelberg, 2010.
2. A. M. Foley, P. G. Leahy, A. Marvuglia, and E. J. McKeogh, "Current methods and advances in forecasting of wind power generation," *Renewable Energy*, vol. 37, no. 1, pp. 1 – 8, 2012.
3. J. de Jess Rubio, "Analytic neural network model of a wind turbine," *Soft Computing*, pp. 1–9, 2014.
4. S. Soman, H. Zareipour, O. Malik, and P. Mandal, "A review of wind power and wind speed forecasting methods with different time horizons," in *North American Power Symposium (NAPS), 2010*, pp. 1–8, Sept 2010.
5. C. Monteiro, H. Keko, R. Bessa, V. Miranda, A. Botterud, J. Wang, and G. Conzelmann, "A quick guide to wind power forecating : state-of-the-art 2009," technical report, Argonne National Laboratory, 2009.
6. X. Wang, P. Guo, and X. Huang, "A review of wind power forecasting models," *Energy Procedia*, vol. 12, no. 0, pp. 770 – 778, 2011. The Proceedings of International Conference on Smart Grid and Clean Energy Technologies (ICSGCE 2011.
7. C. Monteiro, R. Bessa, V. Miranda, A. Botterud, J. Wang, and G. Conzelmann, "Wind power forecasting: State-of-the-art 2009," technical report, Argonne National Laboratory, 2009.
8. C. Kamath and Y. J. Fan, "Finding motifs in wind generation time series data," in *Machine Learning and Applications (ICMLA), 2012 11th International Conference on*, vol. 2, pp. 481–486, Dec 2012.
9. B. Chiu, E. Keogh, and S. Lonardi, "Probabilistic discovery of time series motifs," pp. 493–498, 2003.
10. K. A. Larson and K. Westrick, "Short-term wind forecasting using off-site observations," *Wind Energy*, vol. 9, no. 1-2, pp. 55–62, 2006.
11. V. G. Berdugo, C. Chaussin, L. Dubus, G. Hebrail, and V. Leboucher, "Analog method for collaborative very-short-term forecasting of power generation from photovoltaic systems," in *Next Generation Data Mining Summit: Ubiquitous Knowledge Discovery for Energy Management in Smart Grids and Intelligent Machine-to-Machine (M2M) Telematics*, 2011.
12. O. Ohashi and L. Torgo in *ECAI* (L. D. Raedt, C. Bessire, D. Dubois, P. Doherty, P. Frasconi, F. Heintz, and P. J. F. Lucas, eds.), pp. 975–980, IOS Press.
13. S. Pravilovic and A. Appice, "The intelligent forecasting model of time series," *Automation, Control and Intelligent Systems*, vol. 1, pp. 90–98, 2013.
14. J. Lin, E. Keogh, S. Lonardi, and B. Chiu, "A symbolic representation of time series, with implications for streaming algorithms," in *In Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, pp. 2–11, ACM Press, 2003.
15. C. Guestrin, P. Bodik, R. Thibaux, M. Paskin, and S. Madden, "Distributed regression: an efficient framework for modeling sensor network data," in *Information Processing in Sensor Networks, 2004. IPSN 2004. Third International Symposium on*, pp. 1–10, April 2004.
16. G. E. P. Box and G. Jenkins, *Time Series Analysis, Forecasting and Control.* Holden-Day, Incorporated, 1990.