Online Short-term Forecasting of Photovoltaic Energy Production

Aleksandra Rashkovska, Jošt Novljan, Miha Smolnikar, Mihael Mohorčič, Carolina Fortuna
Department of Communication Systems
Jožef Stefan Institute
Jamova cesta 39, Ljubljana, Slovenia
{aleksandra.rashkovska, jost.novljan, miha.smolnikar, miha.mohorcic, carolina.fortuna}@ijs.si

Abstract—**Short-term forecasting of the energy production is one of the key issues in smart homes that tend to achieve efficient balance among the energy production, storage and consumption. In this paper, we first perform an analysis of the features to be used by the most promising short-term forecast model: artificial neural networks. We determine the best performing offline model and then propose an online model that is very close to the offline model in terms of prediction accuracy. The evaluation is performed on a real world data and the resulting system is part of a proof-of-concept application for microgrid management.**

Index Terms— **Forecasting, Neural networks, Photovoltaic systems, Microgrids**

I. INTRODUCTION

One of the key challenges of smart grids is to integrate the increasing number of distributed energy sources and microgrids into a dynamic and efficient macrogrid system [1], where the produced energy is at any time effectively spent and the peak loads are minimized. This problem is often dubbed as demand-response and concerns the electricity suppliers as well as the prosumers (i.e. energy producers and consumers). The trend is to promote self-sustaining prosumer entities, where most of the produced energy is used at the same location, for instance within smart buildings. Such subsystems should incorporate short-term forecasting of the energy production and consumption as well as the state of the electricity grid (tariff) in order to efficiently balance among production, storage and consumption.

Forecasting the produced energy with high accuracy represents one of the key issues in microgrid control, where the photovoltaic (PV) energy sources are dominating the market [1]. In smart grid systems, the forecasts of both the consumption and the production will enable dynamic pricing models as well as proactive control of the macrogrid network. The approaches for forecasting the PV energy production vary in the literature depending on the considered forecast horizon and the forecasted parameters.

With respect to the forecast horizon, statistical methods outperform methods that use physical models for short-term forecasting of up to 6 hours, while physical models are more accurate for longer term forecasting of up to 24 to 36 hours [2], [3]. The short-term forecasts typically use measured weather and PV system data, and satellite and sky imagery observations of clouds, while the long term forecasts use numerical weather prediction (NWP) models [4]-[8]. The best approaches make use of both data and NWP models.

With respect to the forecasted parameter, most of the related work focuses on forecasting the global solar irradiance from which the value of the solar power is then computed [9]-[13]. Further, in order to predict the actual generated electrical energy, the transformation function of the used conversion technology is applied. These approaches have the disadvantage of not taking into account equipment degradation and modifications in the local environment [8]. For instance, when the properties of the transformation modules degrade or the micro-environment changes, the transformation function may introduce significant errors in the final forecasted value. Adaptive techniques that also use locally measured data are able to provide more accurate predictions by implicitly taking into account the effects of possible dust, snow, dry leaves that gather on the panels, etc.

Most of the related work use offline forecasting techniques [9]-[14] which tend to perform slightly better than online techniques [15], but are slower since they process large batches of data. However, the technology enabling intelligent and proactive prosumer scenarios is very likely to be based on online rather than offline learning. In this paper, we propose a system for online short-term forecasting of the electric current generated by a set of PV panels. After collecting and pre-processing locally measured PV and environmental data, we perform a regression analysis between the descriptive (input) and predicted (output) variables to determine the features relevant for the prediction task. We then evaluate the performance of Artificial Neural Networks (ANNs), including an online version, with a number of engineered feature sets, and discuss the differences. We have selected ANNs as the most promising categories of statistical methods that perform

well for short-term forecasting [8]. Our contributions are: (*i*) we predict an actual measured value coming from the PV power plant that takes into account degradations and changes in the micro-environment, (*ii*) we provide a comprehensive evaluation with respect to the input feature set, (*iii*) we also implement and evaluate an online ANN model in contrast to most work that use models trained offline [9]-[14] and (*iv*) we use real-life data and an open source prediction system for the implementation of the proof-of-concept application.

The paper is structured as follows. Section II discusses and analyzes the collected data. Section III presents and evaluates the offline forecasting methods, while Section IV focuses on the online forecasting. Finally, Section V concludes the paper.

II. DATA COLLECTION AND ANALYSIS

In this section, we describe the data collection procedure and perform an analysis of the collected data. Our dataset consists of two parts: locally collected data by sensors mounted on the PV power plant and external data collected from independent sources.

A. Locally Collected Data

The locally collected data is provided by five sensor nodes installed on monocrystalline silicon solar panels forming a PV power plant situated on a business building in Ljubljana, Slovenia [16]. The setup consists of four clusters of panels[1]: two clusters having their panels oriented towards the south, one cluster having its panels oriented towards the east and the last cluster having its panels oriented towards the west. Between each cluster of panels and the corresponding power inverter, a VESNA sensor node[2] is installed and measures (*i*) the produced current [A] of one solar panel within cluster and (*ii*) the temperature of the PN junctions (modeled with the top and bottom solar panel temperature measurement) [°C]. The four nodes are labeled 51-54 as listed in TABLE I and perform measurements every 2 minutes. The measurements are then sent to a server using the available GPRS connection.

The fifth sensor node hosts a weather station measuring the local weather conditions; namely, it samples every 2 minutes the (*i*) air temperature [°C], (*ii*) relative humidity [%], (*iii*) wind speed [m/s] and direction [deg], (*iv*) precipitation [mm], and (*v*) the intensity of solar radiation in the visible and ultraviolet parts of the spectrum [W/m^2].

The measured sensor data is collected in a database located on a remote server. Each measurement is time-stamped and accompanied with metadata including the ID of the node, and its longitude and latitude. This paper characterizes the time period between May 1st, 2013 at 07:29 to November 14th, 2013 at 23:59. The properties of the data from the sensor network in terms of size and missing values are summarized in TABLE I. For the training and evaluation of the offline and online models, average hourly values were calculated for each variable from the dataset resulting in a summarized dataset of 4649 instances. Missing data were compensated using linear interpolation.

TABLE I. LOCALLY COLLECTED DATASETS FROM THE PV SYSTEM

Node No.	Dataset size [MB]	Total instances	Damaged instances [%]	Remark
1	65.2	174762	0.72	Weather station
51	30.3	107248	33.91	Top temp. missing
52	37.5	131183	3.45	/
53	43.5	156341	12.14	Top & bottom temp. missing
54	49.3	172093	1.03	/

From TABLE I it can be seen that the data collected from the weather station (Node 1) is the most consistent and has the lowest amounts of missing or damaged data (0.72%). It also contains the largest amount of measurements. On the other hand, Node 51 has the most damaged data because there was a malfunction on the sensor measuring the top temperature of the PN junction for 30% of the dataset. On Node 53, there was also an outage of both top and bottom PN temperature sensors for 12% of the dataset. The dataset from Node 54 was selected for further analysis because of its minimum missing and damaged data compared to other nodes.

B. Regression Analysis of the Locally Collected Data

The aim of this analysis is to identify the influential system variables that have a statistically significant correlation with the forecasted PV current. For the analysis, average hourly values for each measured parameter were computed and the correlation was investigated in terms of R^2 statistics using Matlab$^©$. The analysis results for the cases with large windows of missing data (such as 24 hours or more) are excluded. Also, note that the R^2 statistic can be negative if the model is not appropriate for the data.

The results for the regression analysis between the PV current and each local descriptive (input) variable are given in TABLE II. As expected and confirmed in previous studies [14], the results indicate that the most correlated parameter with the PV current is the solar radiation. After the solar radiation, the next most correlated variable to the PV current is the air temperature.

TABLE II. REGRESSION ANALYSIS (R^2) OF THE LOCALLY MEASURED SYSTEM PARAMETERS

Solar radiation	Air temperature	Relative humidity	Precipitation
0.8870	0.2436	-0.2427	-0.0771
Wind speed	**Wind direction**	**Bottom module temperature**	**Top module temperature**
0.0934	-0.2045	0.1270	0.2091

C. External Data

By external data, we refer to other data sources that may help with the prediction of power generation such as sun position, weather forecasts, etc.

Previous studies have confirmed that using additional data sources such as sun position at the forecast horizon and

weather forecast can be used to improve the predictions. Sun elevation and azimuth can be computed according to time of the day, and the latitude and longitude of the sensor. Additionally, weather forecast data from online providers who use NWP models can also be used. Such forecasts can be obtained from web services such as Weather.com, Forecast.io and OpenWeatherMap. The first two offer hourly forecasts, while the latter only offers 3-hour forecasts. The forecasted phenomena include temperature, cloud coverage, UV-index, humidity, etc. The accuracy of the forecasts differs from provider to provider and depends on the geographical location for which the forecast is provided.

In this paper, we used forecasted weather parameters (i.e., solar radiation, air temperature and cloud cover) retrieved from 54-hour forecast for each day from the ALADIN NWP model. The data was provided by the Slovenian Environmental Agency[3]. Using forecast data from non-local providers, such as Forecast.io, increased the mean average error in current prediction by 0.2 A. Furthermore, we calculated the sun elevation and azimuth for the positions of the sensors at each hour of the analyzed period.

III. OFFLINE FORECAST MODELS

A. Experimental Design

ANNs are computational models presented as systems of interconnected "neurons" which can compute output values for the predicted (target) variable from input values for the descriptive variables [17]. In this study, the ANN was trained with the Matlab$^{©}$ software and its *Neural Network toolbox*, more precisely, the *Neural Network Fitting Tool* used to solve an input-output fitting problem with a two-layer feed-forward neural network. The *Neural Network Fitting Tool* incorporates the steps to select input and output data, create and train the network, and evaluate its performance using mean square error (MSE) and regression analysis. MSE is the average squared difference between outputs and targets. Regression R values measure the correlation between outputs and targets. An R value of 1 means a close relationship, 0 a random relationship.

A two-layer feed-forward network with sigmoid hidden neurons (number of hidden neurons = 10) and linear output neurons was used. The network was trained with the Levenberg-Marquardt backpropagation algorithm. The dataset was randomly divided for training (70% of the samples), validation (15% of the samples) and testing (15% of the samples). To speed up the learning process, some preprocessing of the input data was done automatically: normalization (transforms input data so that all values fall into the interval [−1, 1]) and duplicates cleaning (removing the rows of the input vector that correspond to input elements that always have the same value). To ensure that an ANN of good accuracy has been found, the network was trained 10 times and the average of the MSE and R are reported as results.

Four ANN models for forecast horizon of 6-hour ahead were investigated (TABLE III). The first model (Model 1) is based only on relevant local PV and environment data as input

_____
[3] http://www.arso.gov.si/

- the hourly average values at time instances t, t-1h, t-2h, t-3h, t-4h and t-5h of the (*i*) solar radiation (SR), (*ii*) air temperature (AT) and (*iii*) PV current (PVC), while the target is the corresponding hourly average value of the PV current at time instant (t + 6h). Next, in order to investigate to which extent simple data about the sun position at the forecast horizon can improve the prediction, we investigated the same forecasting model, but with two additional external inputs: the elevation (ELV) and the azimuth (AZM) at the forecast horizon (Model 2). Afterwards, we investigated two ANN models (Model 3 and Model 4) that, besides the inputs of Model 2, include also forecasted weather parameters: solar radiation, air temperature and cloud cover (CC). Model 3 includes values for the input parameters at time instances t: t-5h, while Model 4 includes only values at time instant t. The comparison between Model 3 and Model 4 will show how the exclusion of the measured history data as input influences the prediction performance.

TABLE III.   INPUT PARAMETERS FOR THE INVESTIGATED OFFLINE ANN PV FORECASTING MODELS

Model No.	Inputs
1	SR(t:t-5h), AT(t:t-5h), PVC(t:t-5h)
2	SR(t:t-5h), AT(t:t-5h), PVC(t:t-5h), ELV(t+6h), AZM(t+6h)
3	SR(t:t-5h), AT(t:t-5h), PVC(t:t-5h), ELV(t+6h), AZM(t+6h), SR(t+6h), AT(t+6h), CC(t+6h)
4	SR(t), AT(t), PVC(t), ELV(t+6h), AZM(t+6h), SR(t+6h), AT(t+6h), CC(t+6h)

PVC – PV Current, SR – Solar Radiation, AT – Air Temperature, CC – Cloud Cover, ELV – Elevation, AZM – Azimuth. The SR, AT and CC at the forecast horizon are outputs from a NWP model.

B. Results and Discussion

The performance results for the ANN models from TABLE III are given in TABLE IV. The reported results are: (*i*) MSE and (*ii*) R for the training, validation and test dataset; and (*iii*) MSE, (*iv*) mean absolute error (MAE), and (*v*) mean absolute normalized percentage error (MANPE) calculated on the entire dataset (excluding and including nights). MANPE is calculated as MAE normalized by the maximal target PV current.

The R values for Model 1 are smaller than 0.9, which does not indicate good correlation. For Model 2, the correlation is improved. The MANPE on the entire dataset for Model 2 compared to Model 1 decreased by 6.04% calculated excluding nights, and by 6.79% calculated including nights. We can conclude that for longer forecast horizon, the inclusion of sun angle information decreases the overall error significantly.

The comparison of the performance results for Model 2 and Model 3 shows that adding weather forecast information at the time horizon improves the prediction a little, but not significantly. This is due to the accuracy of the weather forecast. Namely, for the considered period (May 2013 – November 2013), the MANPE between the measured and the forecasted solar radiation is 10.73%, and between the measured and the forecasted air temperature is 6.31%. However, Model 3 performs the best with MANPE of 8.42%

calculated on the entire data set excluding nights, or 7.67% including nights. Moreover, an interesting result is that removing the measured history data as input (Model 3 vs. Model 4) improves the overall performance on the entire dataset when nights are included.

TABLE IV. PERFORMANCE RESULTS FOR THE OFLINE ANN MODELS DEFINED IN TABLE III

		Model 1	Model 2	Model 3	Model 4
Train.	**MSE**	6.0361	2.5847	2.0380	1.7675
	R	0.8061	0.9223	0.9395	0.9478
Valid.	**MSE**	6.6737	3.2331	2.4507	1.7444
	R	0.7872	0.9017	0.9267	0.9475
Test	**MSE**	6.8040	3.1803	2.6416	1.7719
	R	0.7822	0.9031	0.9194	0.9473
Entire dataset (exc. nights)	**MSE**	6.1650	2.9065	2.3862	2.8383
	MAE [A]	1.7076	1.0348	0.9382	1.0984
	MANPE [%]	15.3284	9.2892	8.4217	9.8596
Entire dataset (inc. nights)	**MSE**	6.1192	2.5801	2.0576	1.7380
	MAE [A]	1.7009	0.9442	0.8546	0.7291
	MANPE [%]	15.2683	8.4759	7.6712	6.5446

MSE - mean squared error, R - regression correlation, MAE - mean absolute error, MANPE - mean absolute normalized percentage error.

IV. ONLINE FORECAST SYSTEM

A. Experimental Design

For the online ANN, we assumed a classic backpropagation ANN with gradient descent. The used online ANN has *tanh* activation function and an output layer using linear activation function. The settings of the ANN are learning rate of 0.05 and momentum of 0.6.

The main difference between the offline and online experimental design is in the choice of the training and evaluation datasets. While for the offline cases this involves a random selection process as described in Section III B, in the case of online learning, the training and evaluation data are sequential. Each incoming data point is used for training and the evaluation is performed for each predicted value. Same as in the offline case, the inputs are normalized to [-1, 1] which is optimal for learning the ANN with a *tanh* activation function. In contrast to the offline learning, the outputs are not normalized since this improved the learning speed. The results are calculated as MAE of predicted current vs. the measured current throughout the entire dataset, excluding the first 6 weeks when the model is still learning. According to the results in Section III, the two best performing online models selected for further experimentation are summarized in TABLE V.

TABLE V. INPUT PARAMETERS FOR THE INVESTIGATED ONLINE ANN PV FORECASTING MODELS

Model	Inputs
O1	SR(t), AT(t), PVC(t), ELV(t+6h), AZM(t+6h), SR(t+6h), AT(t+6h), CC(t+6h)
O2	ELV(t+6h), AZM(t+6h), SR(t+6h), AT(t+6h), CC(t+6h)

PVC – PV Current, SR – Solar Radiation, AT – Air Temperature, CC – Cloud Cover, ELV –Elevation, AZM – Azimuth. The SR, AT and CC at the forecast horizon are outputs from a NWP model.

B. Experimental Results

The first result concerns the topology of the ANN which is smaller than the topology used by the offline methods. In our case, the best performing topology had one hidden layer with 3 neurons (vs. 10 in offline). We also evaluated larger topologies with more neurons and more layers which all performed worse (see TABLE VI). This result is also consistent with the findings in other application areas [15].

TABLE VI. MAE [A] FOR DIFFERENT ANN TOPOLOGIES OF MODEL O1

# neurons	3	6	9	12
1-layer	1.00	1.08	1.12	1.14
2-layer	1.32	1.28	1.34	1.31

The second result concerns the performance of the online models described in TABLE V. The MAE calculated excluding nights and averaged over 10 experimental runs for the online learning models O1 and O2 is 1.00 A and 1.09 A respectively. This is slightly higher MAE value than the one for the best offline model M3 that has MAE of 0.9382 A. However, it is known that offline models usually outperform online models [15]. In the application area investigated in this paper, the difference in performance in favor of the offline version is only 6%. Online learning is appropriate when the training data is produced online and the network needs to adapt during the process while the system runs on constrained devices with limited computing resources. Therefore, a modest sacrifice on the performance side for the benefit of computational time and resources is acceptable, since we need a system that will perform in real-time.

C. Implementation

The implementation of the system[4] used for online forecasting of the PV current production is based on the QMiner[5] tool that features online ANN. The functional blocks comprising the system are depicted in Fig. 1. The locally measured data points are being streamed to the time stamp alignment and resampling block that sends them to a time window summarization block. The summarized data points are then sent to the training and prediction block from where they can be visualized using a web-based interface shown in Fig. 2.

The proof-of-concept is part of a complete microgrid management dashboard depicted in Fig. 2. The dashboard offers a good and quick overview of the energy status of the

---

[4] https://github.com/josthkko/energy-predict
[5] http://qminer.ijs.si/

microgrid, as well as a simple sensor management system. The top right corner shows instant notifications for the user. The chart in the middle shows the measured and predicted PV current. The numbers above the chart are the forecasted PV current and power consumption 6 hours ahead, the cumulative monetary profit for a given month and the momentary accuracy of the prediction system. Below the chart are custom widgets about the weather forecast, locations of the sensor systems, and a ring chart showing the percentage of total electricity produced by each energy source in the microgrid.
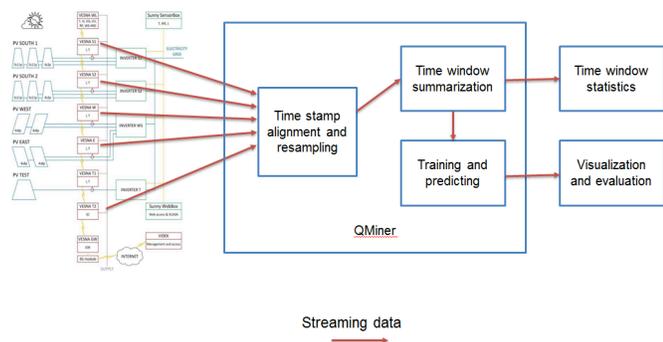

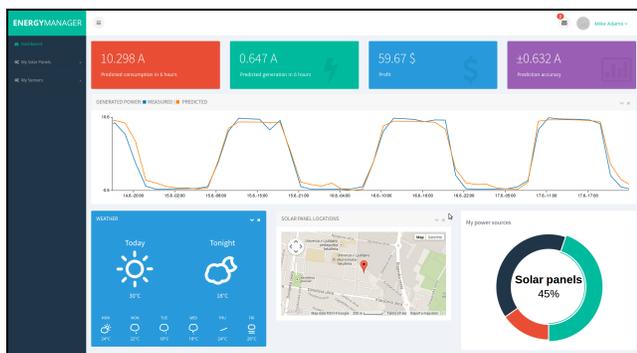Figure 1.   Online PV current production forecast - system overview

Figure 2.   Proof-of-concept microgrid monitoring and control application (http://videk.ijs.si:8080/www/)

V.   CONCLUSION

The paper presents a short-term online forecasting system enabling proactive prosumer scenarios for microgrid developments. The system performs up to 6-hour horizon predictions for the current generated in a real PV power plant located in Ljubljana, Slovenia. The input variables of the models were locally collected sensor network data for past values of hourly PV current production, as well as measured weather variables (solar radiation and air temperature). Additionally, external data for calculated sun angles and forecasted weather parameters (solar radiation, air temperature and cloud cover) at the forecast horizon were included as input. We investigated offline and online ANN models and showed that the performance of the best offline model is just 6% better from the best online one. This result is beneficial for intelligent and proactive prosumer scenarios that are expected to rely on online learning. Moreover, we also present a proof-

of-concept application for microgrid management that integrates online forecasting of PV energy production.

REFERENCES

[1] D. E. Olivares et al., "Trends in Microgrid Control," *IEEE Transactions on Smart Grids*, vol. 5, pp. 1905-1919, July 2014.

[2] Y. Huang et al., "Comparative study of power forecasting methods for PV stations," *Proceedings of POWERCON 2010, International Conference on Power System Technology: Technological innovations making power grid smarter*, Hangzhou, China, 2010, pp. 1-6.

[3] M. Kudo, A. Takeuchi, Y. Nozaki, H. Endo, and S. Jiro, "Forecasting electric power generation in a photovoltaic power system for an energy network," *Electrical Engineering in Japan*, vol. 167, pp. 16-23, Apr. 2009.

[4] J. Kleissl, *Solar Resource Assessment and Forecasting*, Elsevier, 2013.

[5] C. Chow et al., "Intra-hour forecasting with a total sky imager at the UC San Diego solar energy testbed," *Solar Energy*, vol. 85, pp. 2881-2893, Nov. 2011. doi:10.1016/j.solener.2011.08.025

[6] R. Perez, S. Kivalov, A. Zelenka, J. Schlemmer, and K. Hemker, "Improving The Performance of Satellite-to-Irradiance Models using the Satellite's Infrared Sensors," in *Proceedings of ASES Annual Conference*, Phoenix, Arizona, 2010.

[7] H. Pedro and C. Coimbra, "Assessment of forecasting techniques for solar power production with no exogenous inputs," *Solar Energy*, vol. 86, pp. 2017-2028, July 2012.

[8] P. Bacher, H. Madsen, and H. Nielsen, "Online short-term solar power forecasting," *Solar Energy*, vol. 83, pp. 1772-1783, July 2009.

[9] A. Mellit and A. Pavan, "A 24-h forecast of solar irradiance using artificial neural network: Application for performance prediction of a grid-connected PV plant at Trieste, Italy," *Solar Energy*, vol. 84, pp. 807-821, Mar. 2010.

[10] C. Paoli, C. Voyant, M. Muselli, and M. Nivet, "Forecasting of preprocessed daily solar radiation time series using neural networks," *Solar Energy*, vol. 84, pp. 2146-2160, Sept. 2010.

[11] C. Voyant, P. Randimbivololona, M. Nivet, C. Paoli, and M. Muselli, "Twenty four hours ahead global irradiation forecasting using multi-layer perceptron," *Meteorological Applications*, Dec. 2013. doi:10.1002/met.1387

[12] F. Wang, Z. Mi, S. Su, and H. Zhao, "Short-Term Solar Irradiance Forecasting Model Based on Artificial Neural Network Using Statistical Feature Parameters," *Energies*, vol. 5, pp. 1355-1370, May 2012. doi:10.3390/en5051355.

[13] R. Marquez and C. Coimbra, "Forecasting of global and direct solar irradiance using stochastic learning methods, ground experiments and the NWS database," *Solar Energy*, vol. 85, pp. 746-756, Feb. 2011.

[14] M. De Giorgi, P. Congedo, M. Malvoni, and M. Tarantino, "Short-term power forecasting by statistical methods for photovoltaic plants in South Italy," in *Proceedings of the 4th Imeko TC19 Symposium on Environmental Instrumentation and Measurements*, Lecce, Italy, 2013, pp. 171-175.

[15] V. R. Puttige and Sr. G. Anavatti, "Comparison of Real-time Online and Offline Neural Network Models for a UAV," in Proceedings of the International Joint Conference on Neural Networks, Orlando, Florida, USA, August 12-17, 2007, pp. 412 - 417.

[16] M. Mohorcic, M. Smolnikar, and T. Javornik, "Wireless Sensor Network Based Infrastructure for Experimentally Driven Research," in *Proceedings of the Tenth International Symposium on Wireless Communication Systems*, Ilmenau, Germany, 2013, pp. 375-379.

[17] M. Hagan, H. Demuth, and M. Beale, *Neural Network Design*, Boston, MA: PWS Publishing, 1996.