APPLICATION AND EVALUATION OF A PROBABILISTIC FORECASTING MODEL FOR EXPECTED LOCAL PV PENETRATION LEVELS

Raoul BERNARDS
TU Eindhoven
The Netherlands
r.bernards@tue.nl

Ruben VERWEIJ
Stedin BV
The Netherlands

Edward COSTER
Stedin BV
The Netherlands

Johan MORREN
Enexis BV &
TU Eindhoven
The Netherlands

Han SLOOTWEG
Enexis BV &
TU Eindhoven
The Netherlands

ABSTRACT

A data driven scenario based approach is applied to predict the adoption and expected local penetration levels of PV installations in an actual distribution network area in the Netherlands. Local PV adoption probabilities are scaled according to a trained statistical model. Integration of this model in the scenarios is shown to provide a significant improvement in prediction accuracy. Additionally a probabilistic forecast is simulated highlighting the local impact on the electricity network for several future scenarios.

INTRODUCTION

Increasing penetration levels of photovoltaics (PV), electric vehicles and heat pumps are leading to an increased diversity and uncertainty in future residential load profiles [1]. This complicates reliable network planning using traditional load forecasting approaches as large variations in local adoption rates of these new technologies exist. Most long term studies make use of general or aggregated scenarios which are not sufficient to properly assess the local effects of these changes.
In order to better predict the local variation in expected penetration levels of new energy technologies a data-driven scenario based approach was proposed in [2]. A logistic regression analysis was used to correlate households' probability of adoption of a PV system to local geographical, demographic and socio-economic characteristics. Several predictors were found to be statistically significant such as average income level, household composition and age of residents. The developed model can be used to transform a generalized scenario into local specific scenarios, with diversified adoption probabilities per household.
In this paper the predictive power of the model will be evaluated and the proposed scenario-based approach will be applied to predict adoption and expected local penetration levels of PV installations in an actual distribution network area in the Netherlands.

VALIDATION METHOD

In order to validate the improved prediction accuracy obtained by integration of the statistical model, the model is tested on the historical distributed growth of PV over a 10 year period during the years 2006 – 2016 in the network area of Dutch distribution grid operator Stedin.

Scenario characterization

First, a general scenario is determined based on the observed historical growth. The location, size and year of installation of PV were obtained from the Production Installation Register (PIR), which is a database containing information on local generation units in the Netherlands. The growth in the total number of PV installations in the Stedin area is shown in Fig. 1. The general scenario is defined as the exact amount of yearly growth in PV installations as it occurred during the timespan of January 1st 2006 until January 1st 2016.

Generation of households

Individual households in the network area of interest are gathered in a database, where they are coupled with their relevant characteristics. An initial individual adoption probability per household is determined by applying the logistic regression model to the specific characteristics of each household. The complete set of households is then

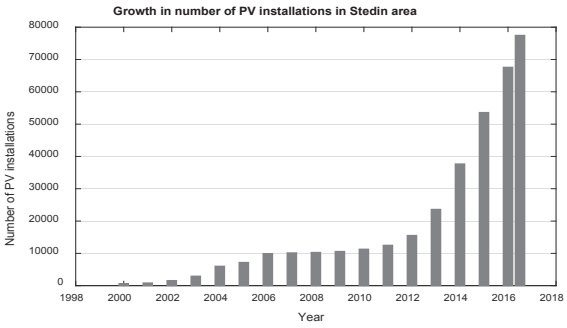

Fig. 1 Number of PV installations in the Stedin area at January 1st of each year.

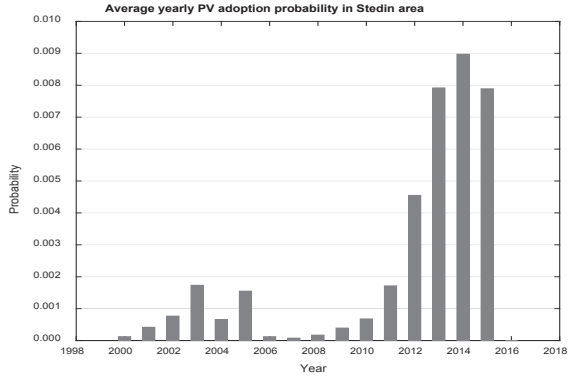

Fig. 2 Average yearly adoption probabilities per customer.

imported into the scenario algorithm, where each household is assigned a unique identifier, a neighbourhood to which it belongs and a local adoption probability. If the household already has a PV installation prior to the start year of the scenario, this is added to the household before the start of the simulation.

Individual adoption probabilities

In order to capture the uncertainty present in the predictions the scenarios are analysed stochastically. The yearly growth of PV installations is converted to a yearly adoption probability by a division by the number of feasible installation locations:

$$p_{scen,n} = \frac{I_n}{C_n - \sum_{i=0}^{n-1} I_i}$$

with I_n the number of PV installations installed in year n, and C_n the number of customers in year n. The term in the denominator (the number of feasible locations) is the total amount of customers minus the total number of locations that already own a PV installation at the beginning of year n. The resulting yearly adoption probabilities are shown in Fig. 2.

Adoption probabilities per household are then scaled according to the chosen scenario. The adoption probability is set independent of whether a household already has a PV installation or not, as this is checked during iteration. Three scenarios are defined using different adoption probabilities:

A. General scenario (no integration of local dependency)
B. Local scenario scaling with relative adoption probability
C. Local scenario scaling with relative log-odds ratio

Simulation A uses just the average yearly adoption probabilities per customer as determined above, so the probability of adoption for house j in year n is:

$$p_{j,n} = p_{scen,n}$$

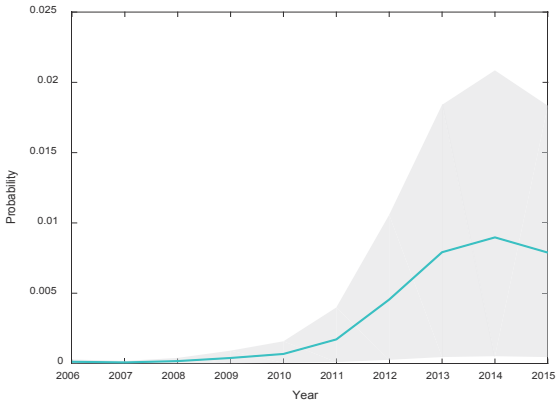

Fig. 3 Spread in predicted adoption probabilities aggregated per neighbourhood.

In simulation B the individual adoption probabilities are transformed to a relative adoption probability with reference to the mean probability and subsequently scaled to align the mean with average scenario probability.

$$p_{j,n} = \frac{p_{adopt,j}}{p_{adopt,mean}} \cdot p_{scen,n}$$

With $p_{adopt,j}$ the adoption probability for customer j determined by the regression model prediction, and $p_{adopt,mean}$ the mean probability of adoption for all customers. In this way the local dependency is integrated while ensuring that the overall predicted expected number of PV installations aligns with the aggregated scenario. Fig. 3 shows the spread in predicted adoption probabilities aggregated per neighbourhood, scaled to match the scenario.

In simulation C the adoption probabilities are determined based on their relative log-odds ratio to the mean. The logistic regression model fits the most likely linear line through the log-odds ratios of the individual households. By relating the adoption probabilities in this way, this relation is respected.

$$logit(p) = \log\left(\frac{p}{1-p}\right)$$

$$logit(p_{j,n}) = logit(p_{scen,n}) + logit(p_{adopt,j}) - logit(p_{adopt,mean})$$

Probabilistic scenario simulation

After determination of the initial household adoptions probabilities the scenario simulations are carried out over the historical 10 year horizon, taking the historical growth as baseline scenario. The flow diagram of the methodology is shown in Fig. 4. A Monte Carlo simulation is executed using 1000 iterations, per iteration 10 years are evaluated where in each year households adopt or do not adopt PV based on a comparison of their individual yearly adoption probability in that year to a generated random number. At the end of each iteration the number of installed PV is aggregated per neighbourhood and written to an output file. Then households are reset to their initial values at start year of the scenario.

For simulation B and C (the local scenarios) the individual adoption probabilities are recalculated during the iteration for each sequential year as these depend on the previous actual adoptions in that iteration.

EVALUATION METRICS

Several performance metrics are used to evaluate the prediction error and accuracy of the different simulation results. The metrics applied for assessing the errors are the Mean Absolute Percentage Error (MAPE), Root Mean Square Error (RMSE) and Continuous Rank Probability Score (CRPS) [3]. The R-squared (R^2) statistic [4] is used as a measure for the prediction accuracy. These metrics are calculated as follows.

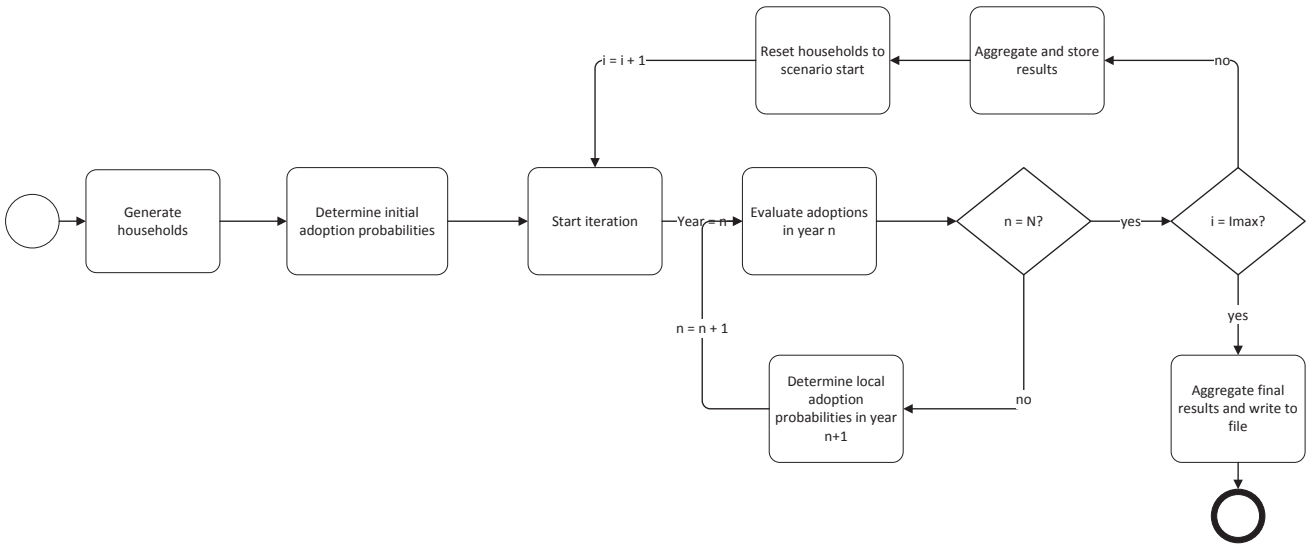

Fig. 4 Flow diagram of the probabilistic scenario methodology.

MAPE

$$MAPE = 100 \cdot \frac{1}{H} \sum_{h=1}^{H} \left| \frac{A_h - F_h}{A_h} \right|$$

With F_h the forecasted value and A_h the actual value of the number of PV installations per neighbourhood h for each of total H neighbourhoods. The advantage of the MAPE metric is that it gives a relatively clearly interpretable value

RMSE

$$RMSE = \sqrt{\frac{1}{H} \sum_{h=1}^{H} (A_h - F_h)^2}$$

The RMSE adds an extra penalty to larger errors due to the square. This is useful to weigh forecasts that are far off more heavily.

CRPS

The CRPS is calculated as the integral of the difference between the cumulative density functions (CDF's) of the forecasted and observed values.

$$CRPS = \frac{1}{H} \sum_{h=1}^{H} \int_{-\infty}^{\infty} \left(D_{A_h} - D_{F_h}(x) \right)^2 dx$$

With D_{A_h} and D_{F_h} the CDF's of respectively the actual observed value and the forecasted value. The CDF of the forecast is obtained by sorting and summing the probability distribution of the forecasted values. The CDF of the actual value can be represented by the Heaviside (step) function with the step from 0 to 1 at the observed value. The CRPS then represents the total area between both CDF's. An advantage of this is that it compares the full distributions and not just a single value, this gives a more accurate measure of comparison for probabilistic forecasts.

R-squared

The R^2 is a commonly used metric in data analysis methods such as linear regression and provides a measure of how well a certain model explains the variance in the outcome. It is calculated as the square of the sample correlation [4] using

$$R^2 = \frac{\sum_{h=1}^{H} (F_h - \bar{A})^2}{\sum_{h=1}^{H} (A_h - \bar{A})^2}$$

where $\bar{A}$ is the mean of the observed values.

SIMULATION RESULTS

This paragraph contains the results for the different simulations and is set up as follows. First a bar plot shows the distribution of total predicted number of PV installations in the entire area per iteration. Then the bias in mean number of PV installations per neighbourhood is shown, compared to the actual values from the PIR. Finally three different performance metrics are shown, for a comparison of the prediction accuracy of the different simulation approaches.

For all simulations the mean of the predicted number of PV installations aligns closely with the observed number from the scenario to which the prediction is tuned. Fig. 5 shows the distribution of the predicted number of PV installations for scenario A, the mean of the forecast, and the observed number of PV installations in year 10. The forecasted number of PV, mean 67622, prediction interval (PI) [67148, 68102], nearly equals the observed value of 67627. The prediction interval is chosen as the 2.5% and 97.5% quantiles, such that 95% of all forecasted values fall within the interval.

Fig. 6 shows a comparison between the errors in forecasted number of PV installations per neighbourhood for scenario A (in gray) and B (in teal). The bias in forecasted number of installations is lower in B (bias PI [-40.7, 19.8]) compared to PI [-56.6, 62.9] in A, especially a decrease in the larger errors can be observed. Fig. 7 shows both scenario B and C to exhibit a significantly reduced prediction error relative to scenario A. The performance metrics for all three scenarios are given in Table 1, showing the improved (and roughly equal) prediction performance in scenarios B and C.

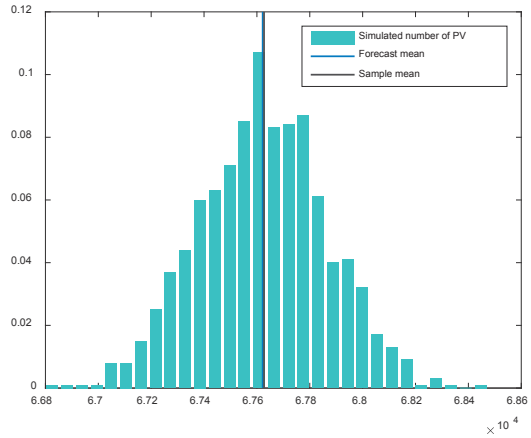

Fig. 5 Distribution of predicted number of PV installations compared to the observed mean.

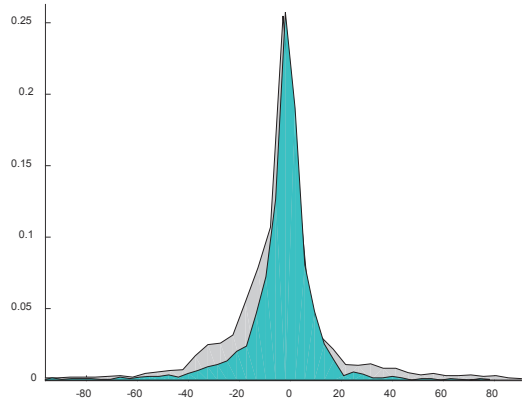

Fig. 6 Comparison of bias in forecasted PV installations per neighbourhood between scenario A and B

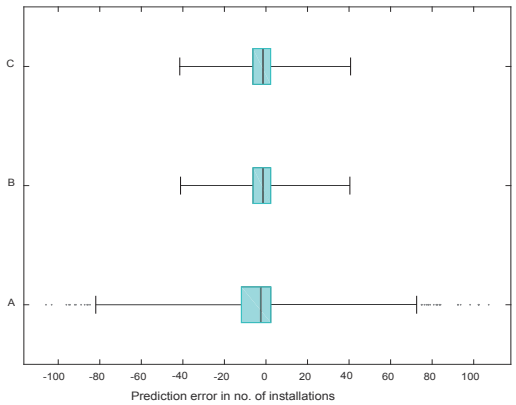

Fig. 7 Spread in prediction error per neighbourhood for each scenario.

Table 1 Evaluation metrics of prediction error and predictive power for different scenarios.

Scenario	MAPE	RMSE	CRPS	R^2
A	101.92%	24.43	5.38	0.44
B	46.37%	15.06	3.88	0.67
C	45.77%	15.02	3.87	0.67

In the next paragraph the diversified adoption probabilities of scenario A and B are used to assess the impact of PV generation on the network.

ASSESSMENT OF FUTURE SCENARIOS

The forecasted number of PV installations for three scenarios are input for investigating the medium voltage (MV) feeder load. Three scenarios are defined which include a 6%, 8% and 11% overall penetration of PV in residential Stedin area in 2020.

PV generation

Two solar profiles were used to calculate PV generation per hour for all feeders. The average profile represents an expected number of sunny hours (around 1700 hours per year in the Netherlands). The extreme solar profile represents a year with more than 2100 sunny hours. The expected installed capacity for new installations is assumed to be 3 kWp. An example of a PV generation profile is shown in Fig. 8.

Impact on feeder load

Multiple feeders were selected to investigate the impact of PV generation on the feeder load. One example is a cable which is feeding a part of the city of Amersfoort. Currently around 600 kWp is installed in this area. According to the simulation results this will increase with 322 kWp in the 6% scenario and 504 kWp in the 8% scenario and 775 in the 11% scenario. The resulting feeder load under the different scenarios is shown in Fig. 9 and Fig. 10 shows the relative load reduction feeder over the year. For the 11% scenario the expected PV reduces the load by more than 20% for approximately 1% of the hours in a year.

The percentage of change in load for three scenarios is given in Table 2.

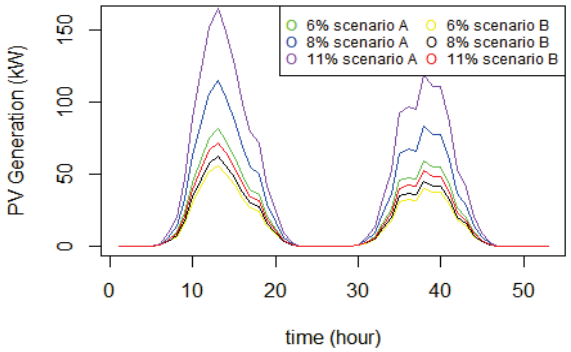

Fig. 8 PV generation on a feeder in Stedin area using the average solar profile.

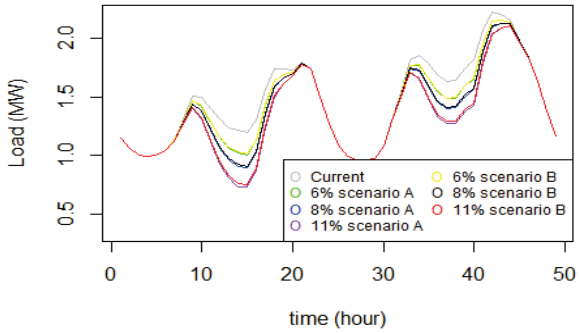

Fig. 9 Impact of PV on feeder load during several winter days for a randomly selected feeder.

Table 2 Summary statistics of load impact due to PV generation in %

	Simulation A			Simulation B		
	6%	8%	11%	6%	8%	11 %
mean	-1.98	-3.12	-4.82	-1.90	-2.98	-4.58
median	-0.032	-0.050	-0.077	-0.030	-0.048	-0.073
min	-17.59	-27.76	-42.94	-16.92	-26.50	-40.76
max	0.00	0.00	0.00	0.00	0.00	0.00

For the investigated feeders the impact of PV generation on the load varied between 0-78% including all scenarios. The percentage varies due to the diversification in PV penetration. Fig. 11 shows the variation in expected installed kWp per feeder between simulation A and the diversified simulation B. Although the mean difference is relatively small, local variations of ±50% in expected installed PV power are not uncommon. The analysis shows the importance of having insight in the development of local penetration of distributed generation to properly estimate the local network impacts of different growth scenarios.

CONCLUSION

Local PV adoption scenarios are generated by scaling individual adoptions probabilities according to a trained statistical model. Using these local scenarios is shown to provide a more accurate representation of the spread of actual expected values than applying a general scenario to an entire area.

Applying the forecasting method to provide a forecast and analysis of a distribution network area for several years into the future showed that the diversification of PV penetration results in substantial variations in local grid impact.

In this way a more accurate estimation of risk levels in specific network areas can be executed as well as a better assessment of the current adequacy of the grid as a whole.

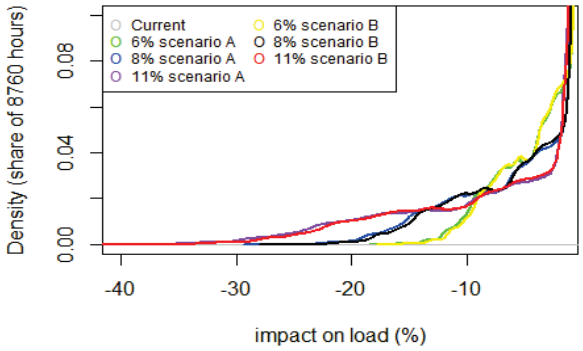

Fig. 10 Density function of percentage of change in feeder load for three scenarios calculated for one year.

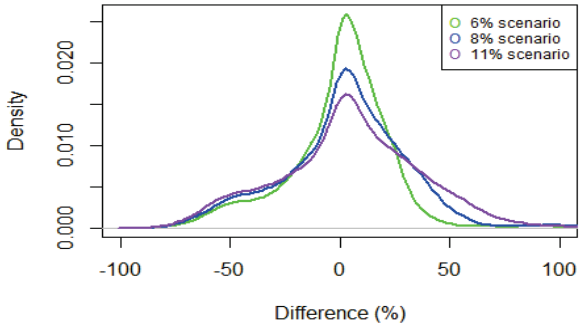

Fig. 11 Variation in expected installed PV power per feeder in percentage.

REFERENCES

[1] M. Hayn, V. Bertsch, and W. Fichtner, 2014, "Electricity load profiles in Europe: The importance of household segmentation", *Energy Res. Soc. Sci.*, vol. 3, 30-45.

[2] R. Bernards, J. Morren and J.G. Slootweg, 2016, "Evaluating Impact of New Technologies on Low Voltage Grids using Probabilistic Data-enriched Scenarios", Proc. 17th EEEIC.

[3] S. Reich and C. Cotter, 2015, *Probabilistic Forecasting and Bayesian Data Assimilation*, Cambridge University Press.

[4] S. Weisberg, 2005, *Applied Linear Regression*, John Wiley & Sons, pp. 352.