University of Cyprus

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

DEGRADATION RATE ESTIMATION IN PHOTOVOLTAICS

ALEXANDER PHINIKARIDES

A Dissertation Submitted to the University of Cyprus
in Partial Fulfilment of the Requirements for the Degree of
Doctor of Philosophy

February, 2017

Approval Page

Doctoral Candidate: Alexander Phinikarides

Doctoral Dissertation Title: Degradation Rate Estimation in Photovoltaics

The present Doctoral Dissertation was submitted in partial fulfilment of the requirements for the Degree of Doctor of Philosophy in the ***Department of Electrical and Computer Engineering*** *and was approved on* ***February 15, 2017*** *by the members of the Examination Committee.*

Examination committee:

Research Supervisor: _____
(Name, position and signature)

Committee Member: _____
(Name, position and signature)

Committee Member: _____
(Name, position and signature)

Committee Member: _____
(Name, position and signature)

Committee Member: _____
(Name, position and signature)

Declaration of Doctoral Candidate

The present doctoral dissertation was submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy at the University of Cyprus. It is a product of original work of my own, unless otherwise mentioned through references, notes, or any other statements.

. . . . . [Full Name]

. . . . . [Signature]

Περίληψη

Η πρόοδος που παρατηρείται τα τελευταία χρόνια στην τεχνολογία κατασκευής φωτοβολταϊκών (ΦΒ) πλαισίων είχε ως αποτέλεσμα την παραγωγή πολύ αποδοτικών ΦΒ τεχνολογιών, γεγονός που συνείσφερε αρκετά στη μείωση του σταθμισμένου κόστους ηλεκτρικής ενέργειας (Levelized Cost of Electricity - LCOE) λόγω της αυξημένης ζήτησης για την τεχνολογία. Δύο βασικοί παράγοντες που θα αυξήσουν τη ζήτηση και θα μειώσουν το LCOE ακόμη παραπάνω είναι: 1) ανάπτυξη του τομέα λειτουργίας και συντήρησης (Operations & Maintenance - O&M) ώστε να εξασφαλίζεται η βέλτιστη λειτουργία των ΦΒ εγκαταστάσεων, και 2) ακριβής εκτίμηση του γνωστού φαινομένου της υποβάθμισης της απόδοσης και η σύγκριση του ρυθμού της υποβάθμισης με την εγγύηση που προσφέρει ο κατασκευαστής των ΦΒ πλαισίων. Και οι δύο αυτοί παράγοντες προϋποθέτουν ενεργή παρακολούθηση και εποπτεία των ΦΒ εγκαταστάσεων, ανάλυση των μετρήσεων που καταγράφονται από το σύστημα παρακολούθησης και εκτίμηση του ρυθμού της μακροχρόνιας υποβάθμισης μέσα στο διάστημα της στατιστικής εμπιστοσύνης. Η ανάλυση αυτή θα επιτρέψει με τη σειρά της τον προγραμματισμό δράσεων για άμβλυνση των αιτιών της χαμηλής απόδοσης και ελαχιστοποίηση της χαμένης ενέργειας. Η ακριβής εκτίμηση του ρυθμού της υποβάθμισης θα επιτρέψει επίσης την ακριβή πρόβλεψη της ενεργειακής απόδοσης της ΦΒ εγκατάστασης καθ' όλη τη διάρκεια της ζωής των ΦΒ πλαισίων. Με αυτό τον τρόπο, θα υπάρχει η δυνατότητα αναθεώρησης των εγγυήσεων που παρέχονται από τους κατασκευαστές έτσι ώστε να παρέχεται πιο αυστηρή εγγύηση, για περαιτέρω μείωση του επενδυτικού ρίσκου και περαιτέρω αύξηση της εμπιστοσύνης προς την τεχνολογία.

Η διατριβή αυτή περιγράφει την ανάπτυξη μιας γενικευμένης μεθολογίας ανάλυσης δεδομένων η οποία στηρίζεται σε στατιστική ανάλυση για την εκτίμηση του ρυθμού της υποβάθμισης κάτω από πραγματικές συνθήκες λειτουργίας, χρησιμοποιώντας πραγματικές μετρήσεις από έντεκα ΦΒ συστήματα διαφόρων τεχνολογιών, τα οποία είναι διασυνδεδεμένα στο δίκτυο. Τα συστήματα αυτά εγκαταστάθηκαν και λειτουργούν από τον Ιούνιο 2006 στο χώρο δοκιμών του Εργαστηρίου ΦΒ Τεχνολογίας του Πανεπιστημίου Κύπρου. Η μεθοδολογία που αναπτύχθηκε, σχεδιάστηκε για να παρέχει ακριβή και εύρωστη εκτίμηση του ρυθμού της υποβάθμισης, χωρίς να χρειάζεται επίβλεψη. Επίσης, σχεδιάστηκε για εφαρμογή σε εμπορικά ΦΒ συστήματα όπου οι δυνατότητες σε αισθητήρες και συστήματα καταγραφής δεδομένων είναι περιορισμένες λόγω κόστους. Έτσι, οι ελάχιστες απαιτήσεις για την εφαρμογή της μεθοδολογίας είναι ακριβείς μετρήσεις της ισχύος που παράγεται από το ΦΒ σύστημα και ακριβής μέτρηση της ηλιακής ακτινοβολίας.

Η μεθοδολογία έχει στόχο να προσφέρει μια πιθανή λύση σε τέσσερα κύρια

προβλήματα στον τομέα της υποβάθμισης των ΦΒ: 1) στην αξιολόγηση των πρωτογενών μετρήσεων και τη δημιουργία του ιδανικού συνόλου των μετρήσεων για περαιτέρω ανάλυση, 2) στην ανίχνευση περιστατικών μη βέλτιστης απόδοσης και αξιολόγηση της επίδρασης τους στον εκτιμημένο ρυθμό της υποβάθμισης, 3) στην ανάλυση χρονοσειρών της απόδοσης και εκτίμηση της γραμμικής ή μη γραμμικής τάσης για υποβάθμιση, και 4) την αντικατάσταση των ειδικών (ad hoc) μεθοδολογιών με στατιστικά κριτήρια για μείωση της μεροληψίας (bias), αυτοματοποίηση της διαδικασίας και γενίκευση της προτεινόμενης μεθοδολογίας. Για να γίνει αυτό, αναπτύχθηκε μια προγραμματιστική μεθοδολογία που αποτελείται από επιμέρους λειτουργίες για αξιολόγηση των μετρήσεων, ανίχνευση μη βέλτιστης απόδοσης και μετριασμού της επίδρασης της και μοντελοποίηση του ρυθμού της υποβάθμισης. Επίσης, αξιολογήθηκε το υπολογιστικό κόστος της εφαρμογής της μεθοδολογίας και εξερευνήθηκαν τρόποι για τη μείωσή του. Επιπλέον, πραγματοποιήθηκαν εκτεταμένες πειραματικές μελέτες για την εκτίμηση του ρυθμού της υποβάθμισης των ΦΒ πλαισίων υπό παρακολούθηση, κάτω από πρότυπες συνθήκες δοκιμής (Standard Test Conditions - STC). Οι μελέτες διεξήχθησαν σε περιβάλλον εργαστηρίου, δηλαδή εκτός του τόπου εγκατάστασης των ΦΒ πλαισίων (ex situ), χρησιμοποιώντας εξειδικευμένο εξοπλισμό ακριβείας (ηλιακός προσομοιωτής, διάταξη ηλεκτροφωταύγειας, διάταξη υπέρυθρης θερμογραφίας) για μέτρηση της ισχύος των πλαισίων σε STC και μη-καταστροφικό χαρακτηρισμό της ποιότητας των ΦΒ κυψέλων. Κατά τη συνολική διάρκεια των δοκιμών που περιγράφονται στη διατριβή αυτή, η ποιότητα των μετρήσεων εξασφαλιζόταν μέσω ιχνηλάσιμης βαθμονόμησης και περιοδικών ελέγχων.

Στην περίπτωση των δοκιμών σε STC, τα αποτελέσματα ήταν πιο ιχνηλάσιμα και ακριβή από τα αποτελέσματα της ανάλυσης των εξωτερικών μετρήσεων, όπως ήταν αναμενόμενο. Από την άλλη, οι δοκιμές σε περιβάλλον εργαστηρίου απαιτούσαν αρκετή χειρονακτική εργασία και προσωρινή διακοπή της κανονικής λειτουργίας των ΦΒ συστημάτων. Έτσι, ως αποτέλεσμα του χειρισμού των ΦΒ πλαισίων κατά την απεγκατάσταση, τη μεταφορά στο εργαστήριο και την επανεγκατάσταση, ο κίνδυνος για πρόκληση ζημιάς ήταν μεγάλος. Τα αποτελέσματα από τις δοκιμές σε STC χρησιμοποιήθηκαν ως σημείο αναφοράς για αξιολόγηση της εγκυρότητας της μεθοδολογίας ανάλυσης δεδομένων που αναπτύχθηκε. Με τον τρόπο αυτό, συγκρίθηκε η απόδοση κάτω από πραγματικές συνθήκες λειτουργίας με την απόδοση κάτω από πρότυπες συνθήκες δοκιμής, αποτέλεσμα πολύ σημαντικό για τον τομέα των ΦΒ, καθώς υπάρχουν ελάχιστα παραδείγματα στη βιβλιογραφία με μακροπρόθεσμες συγκριτικές μελέτες από διάφορες ΦΒ τεχνολογίες.

Abstract

Recent advances in photovoltaic (PV) module manufacturing have resulted in the production of highly efficient cells and modules and the significant reduction of the levelized cost of electricity (LCOE) due to the increased demand for the technology. Two key factors that will increase the demand and reduce the LCOE even further are: 1) improving operations and maintenance (O&M) to ensure the optimal operation of deployed PV plants, and 2) accurately estimating the well-known effect of gradual performance degradation and guaranteeing their lifetime energy yield. Both of these key factors require active monitoring and supervision of the deployed PV plants, analysis of field measurement data for estimation of the long-term degradation rate, R_D with statistical confidence and comparison with the warranty. This analysis will in turn enable the planning of actions to mitigate the causes of low performance and minimize the amount of energy lost. The accurate estimation of the R_D for a deployed PV plant will also enable more accurate and precise lifetime energy yield forecasting and stricter performance guarantees, further reducing investment risk and increasing confidence in the technology.

This work deals with developing a generalized data analysis methodology based on statistical principles, for estimating the energy degradation rate, using field measurement data from eleven different grid-connected PV plants operating side-by-side since June 2006 at the PV Technology test site of the University of Cyprus. The methodology was designed to provide accurate and robust unsupervised estimation with a measure of uncertainty. Also, it was designed for application on commercial PV plants, where sensor deployment is sparse and data logging capabilities are low due to cost. Therefore, the minimum requirements for the realization of the developed methodology are accurate measurements of power and an accurate measurement of irradiance.

The methodology was developed to address four main issues in the field of PV degradation: 1) measurement qualification and creation of a clean data set from uncertain sources, 2) detection of suboptimal performance from the measurement data and assessment of the effect on the actual degradation rate, 3) analysis of time series of constructed performance metrics to extract and analyse the trend in either a linear or non-linear fashion, and 4) substitution of ad hoc analyses and empirical parametrisation with formal statistical tests, to enable the applicability of the methodology in an unsupervised way. Therefore, a data pipeline consisting of measurement qualification, creation of performance metrics, detection and treatment of outliers and trend modelling procedures was developed. In addition, the computational expense of implementing such a methodology was explored and alternative ways were proposed to further reduce it.

Extensive experimental work has also been performed in order to estimate the R_D of the studied PV arrays, under standard test conditions (STC). These ex-

periments were performed ex situ, using high quality laboratory equipment (i.e. flasher, electroluminescence (EL), infrared (IR) thermography), with traceable calibration throughout the evaluation period. Even though the results were more traceable and certain, indoor testing required a significant amount of manual labour and system downtime and introduced risk due to mounting/dismounting and transporting the PV modules to the laboratory. The PV modules under study were characterized in the laboratory and the results were used to benchmark the accuracy of the developed unsupervised methodology. In this way, PV performance measured under a broad spectrum of prevailing meteorological conditions was compared to the results of indoor testing under international standards. This was one of the most important outcomes of this work as this kind of long-term comparison on multiple, co-located PV technologies which were monitored and characterized in a research-grade environment was extremely rare in the bibliography.

Acknowledgements

"The journey of a thousand miles begins with a single step."

– Laozi, Tao Te Ching

With this quote, I am reminded of the beginning of this journey which would not have been the same without the professional and personal relationships formed with the many people I would like to thank.

First and foremost, many thanks to Dr. George E. Georghiou for being my advisor and my mentor all these years and for believing in me. He has done everything possible to see that during my time as a PhD student I was provided with everything needed for the successful outcome of this work. Through his encouragement, his positive attitude and his work ethic he has set an example, which reaches far beyond the scope of this dissertation and for this I will be forever indebted.

Secondly, I would like to thank my friend and colleague Dr. George Makrides with whom we spent many years working side-by-side. His solid technical skills and expertise have been instrumental in providing me with a head-start on the subject of outdoor testing of photovoltaics and publication of research results. Thank you George for the genuine interest, encouragement, support and friendship and the example you set with your work ethic, commitment to quality and endurance.

Special thanks to Dr. Andreas Kyprianou for providing his expertise in signal processing. Without his knowledge and our discussions, many interesting ideas would not have been explored. His level of expertise, his intuition and his overall approach to signal processing have made the work done in the last two years much more interesting.

I am also deeply grateful to the two people whose help was instrumental in the experimental part of this work: Michalis Papastavrou and Marios Tomazou. Many thanks as well to my colleagues Dr. Minas Patsalides, Constantinos Lazarou, Dr. Vaso Paraskeva, Ioannis Koumparou, Nicholas Philippou, Dr. Demetres Evagorou, Dr. Matthew Norton and Despo Demetroude of the UCY for their friendship and the moments we shared all these years.

I would also like to thank Prof. Jürgen H. Werner, Dr.-Ing. Markus Schubert and all colleagues at the Institut für Photovoltaik of the University of Stuttgart for the fruitful discussions and collaboration. Acknowledgements also go to Dr. Marcus Rennhofer and Martin Halwachs of the AIT, Dr. Mauro Pravettoni who was at SUPSI, Dr. Giorgio Belluardo of EURAC, Dr. Clifford W. Hansen of Sandia National Laboratories, Steve Ransome of SRCL and Roberto Galleano and Willem Zaaiman of the Joint Research Center (JRC) for generously sharing their knowledge with me.

The financial support of the Cyprus Research Promotion Foundation for funding part of this work under contract ΤΕΧΝΟΛΟΓΙΑ/ΕΝΕΡΓ/0311(ΒΙΕ)/12 is also appreciated.

Last, but not least, I thank my family and my friends for their understanding, support and confidence in me.

To the memory of my beloved grandmother and my beloved uncle.

Contents

List of Figures

List of Tables

List of Publications

Journal papers

1. M. Hadjipanayi, I. Koumparou, N. Philippou, V. Paraskeva, A. Phinikarides, G. Makrides, V. Efthymiou, and G. E. Georghiou, "Prospects of photovoltaics in southern European, Mediterranean and Middle East regions," *Renewable Energy*, vol. 92, pp. 58–74, Jul. 2016. DOI: 10.1016/j.renene.2016.01.096
2. A. Kyprianou, A. Phinikarides, G. Makrides, and G. E. Georghiou, "Definition and Computation of the Degradation Rates of Photovoltaic Systems of Different Technologies With Robust Principal Component Analysis," *IEEE Journal of Photovoltaics*, vol. 5, no. 6, pp. 1698–1705, Nov. 2015. DOI: 10.1109/JPHOTOV.2015.2478065
3. A. Phinikarides, G. Makrides, B. Zinsser, M. Schubert, and G. E. Georghiou, "Analysis of photovoltaic system performance time series: Seasonality and performance loss," *Renewable Energy*, vol. 77, pp. 51–63, May 2015. DOI: 10.1016/j.renene.2014.11.091
4. A. Phinikarides, N. Kindyni, G. Makrides, and G. E. Georghiou, "Review of photovoltaic degradation rate methodologies," *Renewable and Sustainable Energy Reviews*, vol. 40, pp. 143–152, Dec. 2014. DOI: 10.1016/j.rser.2014.07.155
5. G. Makrides, B. Zinsser, A. Phinikarides, M. Schubert, and G. E. Georghiou, "Temperature and thermal annealing effects on different photovoltaic technologies," *Renewable Energy*, vol. 43, pp. 407–417, Jul. 2012. DOI: 10.1016/j.renene.2011.11.046

Conference papers

1. A. Phinikarides, G. Makrides, and G. E. Georghiou, "Estimation of the Degradation Rate of Fielded Photovoltaic Arrays in the Presence of Measurement Outages," in *32nd EU-PVSEC*, Munich, Germany, 2016, pp. 1754–1757, ISBN: 3-936338-41-8. DOI: 10.4229/EUPVSEC20162016-5DO.12.6
2. A. Phinikarides, C. Shimitra, R. Bourgeon, I. Koumparou, G. Makrides, and G. E. Georghiou, "Development of a Novel Web Application for Automatic Photovoltaic System Performance Analysis and Fault Identification," in *43rd IEEE PVSC*, Portland, OR, USA, 2016, pp. 1736–1740. DOI: 10.1109/PVSC.2016.7749921
3. A. Kyprianou, A. Phinikarides, G. Makrides, and G. E. Georghiou, "Forecasting Degradation Rates of Different Photovoltaic Systems Using Robust Principal Component Analysis and ARIMA," in *32nd EU-PVSEC*, Munich, Germany, 2016, pp. 2033–2035, ISBN: 3-936338-41-8. DOI: 10.4229/EUPVSEC20162016-5BV.2.58
4. I. Koumparou, A. Phinikarides, G. Makrides, and G. E. Georghiou, "Characterisation and Mapping of Daily Sky Conditions Based on Ground Measurements of Solar Irra-

diance in Mainland USA," in *43rd IEEE PVSC*, Portland, OR, USA, 2016, pp. 0986–0991. DOI: 10.1109/PVSC.2016.7749758

5. G. Makrides, A. Phinikarides, E. Herzog, M. B. Strobel, and G. E. Georghiou, "Outdoor Performance and Modelling of Innovative Crystalline Silicon Photovoltaic Modules Under Hot Climatic Conditions," in *32nd EU-PVSEC*, Munich, Germany, 2016, pp. 1991–1996, ISBN: 3-936338-41-8. DOI: 10.4229/EUPVSEC20162016-5BV.2.44
6. G. Makrides, A. Phinikarides, J. Sutterlueti, S. Ransome, and G. E. Georghiou, "Advanced Performance Monitoring System for Improved Reliability and Optimized Levelized Cost of Electricity," in *32nd EU-PVSEC*, Munich, Germany, 2016, pp. 1973–1977, ISBN: 3-936338-41-8. DOI: 10.4229/EUPVSEC20162016-5BV.2.38
7. **[Best Student Paper Award]** A. Phinikarides, G. Makrides, and G. E. Georghiou, "Estimation of annual performance loss rates of grid-connected photovoltaic systems using time series analysis and validation through indoor testing at standard test conditions," in *42nd IEEE PVSC*, New Orleans, LA, 2015, pp. 1–5. DOI: 10.1109/PVSC.2015.7355940
8. A. Phinikarides, G. Makrides, and G. E. Georghiou, "Analysis of the field performance of a double junction amorphous silicon photovoltaic module and its correlation to standardized testing," in *31st EU-PVSEC*, Hamburg, Germany, 2015, pp. 1992–1996. DOI: 10.4229/EUPVSEC20152015-5AV.6.20
9. **[Best Poster Award]** A. Kyprianou, A. Phinikarides, G. Makrides, and G. E. Georghiou, "Robust Principal Component Analysis For Computing The Degradation Rates Of Different Photovoltaic Systems," in *29th EU-PVSEC*, Amsterdam, 2014, pp. 2939–2942. DOI: 10.4229/EUPVSEC20142014-5BV.2.41
10. **[Finalist for Best Poster Award]** A. Phinikarides, G. Makrides, N. Kindyni, and G. E. Georghiou, "Comparison of trend extraction methods for calculating performance loss rates of different photovoltaic technologies," in *40th IEEE PVSC*, Denver, CO, 2014, pp. 3211–3215. DOI: 10.1109/PVSC.2014.6925619
11. A. Phinikarides, N. Philippou, G. Makrides, and G. E. Georghiou, "Performance loss rates of different photovoltaic technologies after eight years of operation under warm climate conditions," in *29th EU-PVSEC*, Amsterdam, 2014, pp. 2664–2668. DOI: 10.4229/EUPVSEC20142014-5BV.1.27
12. A. Phinikarides, G. Makrides, and G. E. Georghiou, "Comparison of analysis methods for the calculation of degradation rates of different photovoltaic technologies," in *28th EU-PVSEC*, Paris, France, 2013, pp. 3973–3976. DOI: 10.4229/28thEUPVSEC2013-5BV.4.39
13. A. Phinikarides, G. Makrides, N. Kindyni, A. Kyprianou, and G. E. Georghiou, "ARIMA modeling of the performance of different photovoltaic technologies," in *39th IEEE PVSC*, Tampa, FL, Jun. 2013, pp. 797–801, ISBN: 978-1-4799-3299-3. DOI: 10.1109/PVSC.2013.6744268
14. G. Makrides, A. Phinikarides, and G. E. Georghiou, "Performance loss rates of grid-connected photovoltaic technologies in warm climates," in *Global Conference On Renewables and Energy Efficiency for Desert Regions*, 2013, pp. 300–304
15. A. Phinikarides, G. Makrides, and G. E. Georghiou, "Initial performance degradation of an a-Si/a-Si tandem PV array," in *27th EU-PVSEC*, Frankfurt, Germany, 2012, pp. 3267–3270. DOI: 10.4229/27thEUPVSEC2012-4BV.2.16
16. G. Makrides, B. Zinsser, A. Phinikarides, M. Schubert, and G. E. Georghiou, "Temperature and thermal annealing effects on amorphous silicon PV," in *26th EU-PVSEC*,

Hamburg, Germany, 2011, pp. 3600–3603. DOI: 10.4229/26thEUPVSEC2011-4AV.2.38

17. G. Makrides, B. Zinsser, A. Phinikarides, M. Norton, G. E. Georghiou, M. Schubert, and J. H. Werner, "Photovoltaic model uncertainties based on field measurements," in *37th IEEE PVSC*, Seattle, WA, Jun. 2011, pp. 2386–2390, ISBN: 978-1-4244-9965-6. DOI: 10.1109/PVSC.2011.6186430
18. A. Phinikarides, G. Makrides, and G. E. Georghiou, "A comprehensive methodology for outdoor and indoor degradation studies on photovoltaic modules," in *3rd International Conference on Renewable Energy Sources & Energy Efficiency*, Nicosia, Cyprus, 2011, pp. 85–93
19. M. Norton, A. Dobbin, A. Phinikarides, T. Tibbits, G. E. Georghiou, and S. Chonavel, "Field performance evaluation and modelling of spectrally tuned quantum-well solar cells," in *37th IEEE PVSC*, Seattle, WA, Jun. 2011, pp. 534–537, ISBN: 978-1-4244-9965-6. DOI: 10.1109/PVSC.2011.6186011

Acronyms

AC	alternating current
ACF	autocorrelation function
ADF	augmented Dickey–Fuller
AIC	Akaike information criterion
AICc	corrected Akaike information criterion
AIT	Austrian Institute of Technology
ALM	augmented Lagrange multipler
ANN	artificial neural network
ANOVA	analysis of variance
AO	additive outlier
AR	auto-regressive
ARIMA	autoregressive integrated moving average
a-Si	amorphous silicon
AUR	Arch User Repository
BIC	Bayesian information criterion
BLAS	Basic Linear Algebra Subprograms
BOS	balance-of-systems
CapEx	capital expenses
CDF	cumulative distribution function
CdTe	cadmium telluride
CI	confidence interval
CIGS	copper indium gallium (di)selenide
CPU	central processing unit
CSD	classical seasonal decomposition
c-Si	crystalline silicon
DC	direct current
DFT	discrete Fourier Transform
DLIT	dark lock-in thermography
EL	electroluminescence
EVA	ethylene-vinyl acetate
FFT	Fast Fourier Transform
FIR	finite impulse response
flop	floating-point operation
GPL	General Public License

GWN	Gaussian white noise
HIT	heterojunction with intrinsic thin-layer
HW	Holt-Winters exponential smoothing
IALM	inexact augmented Lagrange multiplier
IEC	International Electrotechnical Commission
i.i.d.	independent and identically distributed
IIR	infinite impulse response
ipv	Institut für Photovoltaik, Universität Stuttgart
IQR	interquartile range
IR	infrared
IRENA	International Renewable Energy Agency
IV	current-voltage
JRC	Joint Research Centre
KPI	key performance indicator
KPSS	Kwiatkowski–Phillips–Schmidt–Shin
LAPSS	large area pulsed solar simulator
LCOE	levelized cost of electricity
LI	linear interpolation
LOCF	last observation carried forward
LOESS	locally weighted smoothing (LOcal regrESSion)
LR	linear regression
LS	level shift
MA	moving average
MAE	mean absolute error
MAPD	mean absolute percentage deviation
MAPE	mean absolute percentage error
MCAR	missing completely at random
MI	multiple imputation
MKL	Math Kernel Library
MLE	maximum likelihood estimation
MMF	mismatch factor
MO	multiple overimputation
mono-Si	monocrystalline silicon
MPP	maximum power point
MSE	mean squared error
μc-Si	microcrystalline silicon
NREL	National Renewable Energy Laboratory
O&M	operations and maintenance
OLS	ordinary least-squares
OpEx	operational expenses
PACF	partial autocorrelation function

PCA	principal component analysis
PCP	principal components pursuit
PDF	probability density function
PE	percentage error
PID	potential induced degradation
POA	plane of array
poly-Si	polycrystalline silicon
PSD	power spectral density
PTC	PVUSA test conditions
PV	photovoltaic
PVUSA	Photovoltaics for Utility-Scale Applications
RAM	random access memory
regARIMA	regression model with ARIMA errors
RMSE	root mean squared error
RNG	random number generator
RPCA	robust principal component analysis
rRPCA	randomized robust principal component analysis
RSS	residual sum of squares
rSVD	randomized singular value decomposition
SEATS	signal extraction in ARIMA time series
SI	single imputation
STC	standard test conditions
STL	seasonal-trend decomposition by LOESS
SVD	singular value decomposition
SWE	Stæbler-Wronski effect
TCO	transparent conductive oxide
TRAMO	time series regression with ARIMA noise, missing values and outliers
TS	Theil-Sen
UCY	University of Cyprus
UV	ultraviolet
X-12-ARIMA	seasonal adjustment technique developed by the U.S. Census Bureau
X-13ARIMA-SEATS	update of X-12-ARIMA with SEATS-TRAMO
Y-o-Y	year-over-year
ZSW	Zentrum für Sonnenenergie- und Wasserstoff-Forschung

Symbols

Symbol	Description	Unit
θ_{AOI}	angle of incidence	$^\circ$
η_{STC}	module efficiency at STC	%
a_W	wind direction	$^\circ$
α_I	temperature coefficient of current	%/K
AM	air mass	a.u.
β_V	temperature coefficient of voltage	%/K
DHI	diffuse horizontal irradiance	$\mathrm{W/m^2}$
DNI	direct normal irradiance	$\mathrm{W/m^2}$
E_A	net energy from array	kW h
E_{TUN}	net energy to utility grid	kW h
FF	fill factor	a.u.
G_I	total irradiance incident on the surface of a PV device	$\mathrm{W/m^2}$
G_{STC}	global irradiance at STC	$1000\,\mathrm{W/m^2}$
γ_P	temperature coefficient of power	%/K
GHI	global horizontal irradiance	$\mathrm{W/m^2}$
H_I	total irradiation incident on the surface of a PV device	$\mathrm{kW\,h/m^2}$
H_{rel}	relative humidity	%
I_A	array current	A
I_{MPP}	current at maximum power point (MPP)	A
I_{SC}	short-circuit current	A
I_{TU}	current to utility grid	A
iPR	instantaneous performance ratio	a.u.
L_c	array capture losses	h
P_0	maximum power of a PV device at STC prior to exposure	W
P_A	direct current (DC) power from a PV array	W
P_{MPP}	DC power at MPP	W
P_{nom}	nameplate capacity of a PV device at STC	W

Symbol	Description	Unit
P_{PTC}	power extrapolated to Photovoltaics for Utility-Scale Applications (PVUSA) test conditions	W
P_{STC}	maximum power of a PV device at STC	W
P_{TU}	alternating current (AC) power to the utility grid	W
PR	performance ratio	a.u.
PR_{TC}	temperature-corrected performance ratio	a.u.
R_D	degradation rate of PV performance	%/y
R_{D_C}	degradation rate of PV power at STC	%/y
R_{D_E}	degradation rate of PV performance in the field	%/y
r_S	series resistance	Ω
r_{SH}	shunt resistance	Ω
S_W	wind speed	m/s
T_{am}	ambient air temperature	˚C
T_m	module temperature	˚C
T_{NOCT}	nominal operating cell temperature	˚C
V_A	array voltage	V
V_{MPP}	voltage at MPP	V
V_{OC}	open-circuit voltage	V
V_{TU}	utility voltage	V
Y_A	array yield	h
Y_f	final yield	h
Y_r	reference yield	h

Notation

∇^d	dth order difference operator
α	significance level
$\mathcal{B}$	back-shift operator
D_n	Kolmogorov–Smirnov test statistic
F_X	cumulative distribution function of random variable X, $F_X(x) = P(X \leq x)$
f_X	probability density function of random variable X, $f_X(x) = \frac{d}{dx}F_X(x)$
H_0	null hypothesis
H_1	alternative hypothesis
#	hash function that maps data of arbitrary size to data of fixed size
$\hat{x}$	mean of x
$\mathbb{N}$	set of all natural numbers
$\Phi(\mathcal{B})$	autoregressive operator
ρ	correlation
$\mathcal{S}$	a set
$\langle \mathcal{S} \rangle_\tau$	average of set $\mathcal{S}$ over the interval defined by τ
σ	standard deviation
$\hat{\theta}$	estimator of parameter θ
W	Shapiro-Wilk test statistic
$\mathbf{x}_{:,j}$	jth column of matrix $\mathbf{X}$
$\mathbf{x}_{i,:}$	ith row of matrix $\mathbf{X}$
$x_{i,j}$	scalar element found in the ith row and jth column of matrix $\mathbf{X}$

Chapter 1

Introduction

1.1 Motivation and background

Estimates of degradation rate, R_D, are essential in assessing the effective lifetime of a photovoltaic (PV) module, given the 25 to 30-year long warranties offered by manufacturers, who guarantee the specified nameplate capacity, P_{nom} with a maximum of 20 % degradation by the end of the service lifetime. In the last few years, manufacturers have extended this guarantee to provide maximum linear degradation rates per year. Typical guarantees state that the PV module's P_{nom} will not degrade higher that 1 %/y in the first ten years of operation. For the rest of their service lifetime, the guarantee states that the P_{nom} will not degrade higher than 0.67 %/y to 0.8 %/y, depending on the manufacturer. The fact that warranties offered by manufacturers are by definition not the product of testing the PV modules to the end of their lifetime in the field, further validates the need for establishing a standardized methodology for accurately estimating degradation rates of fielded PV modules and systems.

Even by disrupting normal plant operation to measure the module's performance under laboratory conditions, at standard test conditions (STC), the measurement uncertainties are high enough that a difference lower than 2 %–3 % from the P_{nom} will be within the experimental uncertainty. Additionally, by solely testing at STC, a very narrow window of field performance is tested, since in most outdoor environments, such conditions rarely occur, if at all. Therefore, the overall performance of the array/plant in the field needs to be analysed in order to draw inference on the actual rate of degradation in the field, i.e. under real operating conditions.

The main motivation behind this dissertation is the lack of a generalized methodology for assessing such an important aspect of PV plant performance, without interrupting normal plant operation. Although much effort has been put into estimating the degradation rate, no methodology has been proposed and aimed towards minimizing the bias generated by the person performing the analysis due to the selection of ad hoc analysis techniques, i.e. measurement filtering, outlier handling, trend modelling and interpretation of the results. It is for this reason that the estimation of the degradation rate of PV remains a highly

controversial topic.

Especially with respect to handling measurement data, current best practices involve discarding a very large amount of intra-daily measurements and focusing on the performance around noon in order to minimize seasonality and secondary effects such as reflections, cloud motion and temperature fluctuations. Additionally, current best practices in handling outliers and missing data involve either ignoring them completely or using previously validated historical data from similar days to fill in the gaps. Finally, the most widely used method in the literature for estimating R_D, utilizes linear regression with ordinary least-squares (OLS) to fit a linear trend on data reduced by the aforementioned filtering techniques. As has been proven numerous times in the literature, a simple OLS trend is not the universal answer. Even more concerning is the universally accepted practice in the field that the aforementioned data reduction and normalization techniques should be designed as such, to reduce the highly seasonal and uncertain measurement data set to approximate a monotonically decreasing linear process.

The work developed in this dissertation aims to show that by leveraging proper statistical analyses, the current arbitrary practices can be replaced by generalized methodologies which can provide inference, eliminate the selection bias and allow the quantification of uncertainty while at the same time extracting as much information as possible from all useful measurement data in an unsupervised manner.

1.2 Problem statement

As of recently, the proliferation of PV system installations around the globe has resulted in the exponential growth of energy produced from renewable energy sources which is estimated to continue increasing monotonically. Also, according to International Renewable Energy Agency (IRENA), solar PV prices have been on a downward trend [1] and this has resulted in a dramatic reduction of the levelized cost of electricity (LCOE), which is defined in Eq. 1.1:

$$LCOE = \frac{CapEx + \sum_{n=1}^{N} \frac{OpEx - RV}{(1+r)^n}}{\sum_{n=1}^{N} \frac{Y_0(1-R_{D_E})^n}{(1+r)^n}} \tag{1.1}$$

where:

- N = PV system lifetime [years],
- $CapEx$ = total capital expenses [€/kW$_p$],
- $OpEx$ = annual operational expenses [€/kW$_p$],
- RV = residual value [€/kW$_p$],
- r = discount rate [%],

- Y_0 = initial yield [kW h],
- R_{D_E} = energy degradation rate [%].

From Eq. 1.1 it can be seen that one of the major factors in the reduction of the LCOE is the energy degradation rate, R_{D_E}. Up to now, no statistically sound and universally accepted methodology has been developed for its estimation.

In this framework, the estimation of the degradation rate of PV presents several issues; one of them is the lack of quality comparisons between field measurement data and measurements under STC. This is important as through these kinds of comparisons, it can be inferred whether a methodology could be used to estimate a statistically significant rate of degradation, or whether it was due to chance. Another problem is the lack of a universally accepted methodology for assessing degradation and the numerous and different methodologies found in the literature which make different assumptions and result in different evaluations. Due to this, there is currently much effort being put into defining and accepting a universally "correct" methodology [2].

In addition, currently published work on degradation has been concerned with defining methodologies that minimize the uncertainty by trying to linearize the time series of the chosen key performance indicator (KPI) in order to fit a straight line through the data, the slope of which would represent the R_{D_E}. This presents several problems on its own, namely:

- Trying to minimize the confidence interval by discarding unfavourable data points biases the results.
- Reducing the amplitude of the seasonal component by linearly normalizing with a single temperature coefficient introduces error, as the $P \propto T_m$ relationship is not perfectly linear throughout the lifetime of a PV module, as has been proven in field tests.
- Current trend estimation methodologies produce single point estimates and do not calculate the uncertainty.
- As many covariates as are available are used to explain the array power, P_A, without accounting for multicollinearity.

Thus, it is evident that more challenging and far stricter quality assessment of PV installations will be needed, to maintain optimal performance and provide increased confidence in the short- and long-term energy production of the enormous amount of PV installations that will be commissioned in the near future. Such strict assurance cannot be achieved by current practices, which are more specific than general and were designed to solve ad hoc problems in the assessment of PV performance. To meet these new requirements of the PV sector, a new paradigm for performance assessment will be needed. At its centre is applied statistical analysis and a significant amount of effort is currently being dedicated to

converting legacy methodologies into generalised statistical analyses. Applications of statistical analyses in PV can be found in modelling, forecasting, supervision, fault detection and estimation of degradation.

1.3 Aim of this work

The dissertation takes on a holistic view to the estimation of performance degradation of PV systems deployed in the field. The aim was to propose a generalized methodology that treated data deficiencies such as uncertainty, outliers and missing data from field measurements and investigate its accuracy by comparing it to standardized indoor procedures, with minimal requirements in sensor and data logging equipment. The methodology was also aimed at minimizing bias generated by ad hoc methods, by relying on established statistical procedures and formal tests.

One additional aim of this work was to improve estimation of the R_{D_E} without relying on labelled data (e.g. log of outage periods, faults etc.) The tedious manual procedures of the past were uncertain and had to be performed offline. Due to this, they were not repeatable and transferable across the field.

Finally, the choice of the measurement datasets used in this work do not constrain the applicability of the proposed method. All methods proposed and discussed in this dissertation were specifically chosen in order to construct a generalized methodology.

1.4 Outline of the dissertation

This dissertation is structured in such a way that it presents a logical flow for solving the degradation estimation problem in PV. It is divided into eight chapters, including this introduction:

- Ch. 1 presents an overview of the problem, motivation and the specific aims of this work.
- Ch. 2 presents a through literature review in relation to the estimation of degradation rates.
- Ch. 3 presents the experimental setup and the procedures for ensuring measurement quality throughout this work.
- Ch. 4 explains how the data was organized to become amenable for analysis.
- Ch. 5 presents the developed methodology for detecting and handling outliers and missing data using various techniques and assesses the uncertainty.
- Ch. 6 presents the proposed methodologies on degradation rate estimation, its linearity, its sensitivity on the choice of the analysis and the total uncertainty.

- Ch. 7 presents the extensive experimental work performed for validating the statistical analysis approach and their inter-comparison.
- Ch. 8 includes a summary of this work, the research achievements and innovations emanating from this thesis and proposes future work directions.

In addition to the main chapters, a short discussion of the computational cost of the approach and challenges faced throughout this work, along with implemented alternative solutions, is included in Appendix A.

Finally, a description of general contributions to the PV and data science community, stemming in part from this work, is included in Appendix B.

Chapter 2

Related Work

Work from this chapter has been published in [3, 4]

2.1 Introduction

The degradation of PV is currently a much researched topic which has resulted in module production improvements, efficiency increases and the introduction of new materials. Performance degradation can be evidenced in all PV devices, i.e. cell, module, array and system with different factors and degradation mechanisms manifesting on each device. In all cases, the main extrinsic factors related to performance degradation in the field include: temperature, humidity, precipitation, dust, snow and solar irradiance. At the array level, all of these and additionally shading and module mismatches contribute to degradation. The aforementioned factors give rise to various degradation mechanisms [5, 6, 7, 8] and impose significant stress over the lifetime of a PV module, resulting in the reduction of durability, which must be quantified through the estimation of the rate of degradation.

More specifically, at the PV cell level the main mechanisms behind performance loss and possible failure are corrosion, light-induced degradation, contact stability and cracked cells [9, 10]. At the module level, degradation occurs due to the reliability issues of the individual cells and in addition due to glass breakage, delamination, busbar failure, broken interconnects, front surface discoloration, moisture ingress, reduced interlayer adhesion, diode failures and hot-spots. The majority of studies on crystalline silicon (c-Si) technology report that the degradation of power at maximum power point (MPP), P_{MPP}, was mainly attributed to short-circuit current, I_{SC}, losses, followed by smaller decreases in fill factor, FF [11, 12, 13]. I_{SC} degradation was associated with the reduction of P_{MPP} and was most commonly caused by delamination and discoloration [14, 15]. Hishikawa, Morita, Sakamoto, and Oshiro [16] have shown that the reduction in I_{SC} was due to discoloration or delamination at the cell/ethylene-vinyl acetate (EVA) interface, front glass breakage and increased series resistance, r_S, due to the degradation in electrode soldering. A study by National Renewable Energy Laboratory (NREL) suggested that the degradation rate and associated I_{SC} decline were caused by ultraviolet (UV) light absorption at or near the top

of the silicon surface, which caused discoloration [17]. Sanchez-Friera et al. [12] attributed the large degradation rate and I_{SC} losses to delamination of the cell-encapsulant interface, oxidation of the front metallization grid and the anti-reflection coating of the cells and front glass soiling. On the other hand, for thin-film technologies, there was a higher degradation rate of the FF in comparison to the c-Si case [18] and additional mechanisms not observed on c-Si technologies [19, 20, 21]. Finally, at the system level, degradation was the result of individual module failures, array shading, potential induced degradation (PID) [22] and other balance-of-systems (BOS) effects such as inverter efficiency loss, interconnect and cabling losses.

The extent by which the various degradation mechanisms affect different PV technologies does not appear to be identical but depends on the technology, the operating topology and the cumulative history of field exposure, which depends on the location of installation [23]. Consequently, the nominal performance of different PV systems/arrays/modules can be found to degrade at different rates. This rate is expressed as R_D, is defined as the rate of nominal performance drop over time and is denoted as a positive quantity. It is commonly expressed in %/y and represents the reduction of the performance metric in the field [24].

Jordan, Kurtz, VanSant, and Newmiller [25] have compiled a comprehensive review of published degradation rates, both from indoor testing and field data analysis. In summary, from 11 029 published studies, spanning every module technology, the mean R_D was calculated at 0.91 %/y, with the majority of studies published for c-Si. Regarding thin-film technologies, 455 published studies have shown that the mean R_D was higher, at 1.38 %/y, with the majority of studies published after 2000. These results did not include initial degradation. Although the mean for all technologies was 0.93 %/y, the values for thin-film technologies were spread from −1 %/y to 6 %/y, whereas for c-Si the rates were mainly concentrated around the mean. This signifies a very large variation in reported degradation rates, which may be attributed to the small number of field studies and the variability in degradation rate estimation methodologies [26, 27].

Degradation estimation has been the target of international research groups [28, 23], research centres [25] and industry [29, 30, 31, 32]. Methods for estimating R_D vary widely with the choice of performance metric, normalization parameter, data filtering and also test conditions (field vs indoors) [33]. Indoor testing at STC using solar simulators is less often used as it is time consuming and inefficient for other than small-scale PV plants [34]. For this reason, only a small sample of PV modules is tested at STC [35], hence assuming that the degradation of the specific sample of modules is representative of the whole PV plant. The current problems faced in the estimation of the R_D of PV systems in the field, reveal that a multi-disciplinary approach must be adopted to address the multiple shortcomings.

2.2 Photovoltaic performance analysis

In the centrepiece of all research around the estimation of degradation is the actual data used in the analysis which are recorded from the field [36]. Most of the current methodologies have not been proven to be robust to the absence of accurate measurement data [37], hence much effort has been put into creating reliable and accessible monitoring and supervision systems for PV [38, 39]. A large part of PV performance analysis is also concerned with the software design, algorithm development and data warehousing [40, 41, 42] which has received numerous developments in the last few years, mainly because of the lowered barriers to entry for data science and machine learning.

To address the fact that PV has become very accessible for small scale deployment, i.e. on individual homes, and the fact that the cost of the majority of PV monitoring systems becomes prohibitive for this scale, easily deployable solutions [43] have been created using commodity hardware [44, 45] to monitor PV production. Other proposed systems provide more advanced functionality, through real-time diagnostics [46] and were designed to scale [47] while others were designed to make use of satellite data to address the absence of expensive sensor hardware on-site [48].

2.2.1 Performance metrics

The estimation of degradation rates relies on the analysis of chronological ratings of the performance of PV in the field and the prevailing meteorological conditions. Typical parameters include: 1) array current, I_A, array voltage, V_A, and subsequently P_A, as a calculated value, 2) power to the utility grid, P_{TU}, for grid-connected systems, 3) I_{SC}, open-circuit voltage, V_{OC}, FF, r_S, and shunt resistance, r_{SH}, extracted from continuous current-voltage (IV) characterization of modules and arrays in the field, and 4) meteorological measurements such as global irradiance, G_I, ambient air temperature, T_{am}, module temperature, T_m, wind speed, S_W, and relative humidity, H_{rel} [36].

These measurements are used to create chronological performance metrics. Common performance metrics can be grouped into four categories, 1) electrical parameters from IV curves recorded under outdoor or controlled indoor conditions and corrected to STC, 2) extrapolated metrics such as power extrapolated to Photovoltaics for Utility-Scale Applications (PVUSA) test conditions, P_{PTC} [49, 50], 3) normalized metrics such as performance ratio, PR, and 4) scaled metrics such as P_A/P_{nom}, P_{TU}/P_{nom} and array yield, Y_A/P_{nom} [51].

IV curve parameters

Outdoor IV curves are usually obtained at fixed intervals, whereas indoor IV curves are obtained at STC at sparse intervals [52], due to the manual work required to obtain them. From an IV curve, degradation can be observed on the individual electrical parameters [53]. As described in Sec. 2.1, degradation of an electrical parameter can be traced back to the existence of physical defects and degradation mechanisms. Further identification of some

degradation mechanisms can also be performed indoors, with techniques such as electroluminescence (EL), dark lock-in thermography (DLIT) and infrared (IR) thermography. IV characterization in the field is currently mostly performed for research [11] and diagnostic purposes [54], with the modules ideally held at MPP between IV scans, in order to simulate the full load condition [55, 56].

Indoor IV characterization is less commonly used as it is time consuming and inefficient for fielded PV arrays. Furthermore, indoor IV characterization carries the risk of damaging the modules due to mishandling, dismounting from the array and transportation. Thus, it is more efficiently used for standalone modules that are deployed alongside a larger PV system [57], whose purpose is to track degradation under real operating conditions. When using indoor IV characterization, the degradation rate can be calculated as the percentage error (PE), between two successive temporal ratings [58].

Extrapolated metrics

Regression models that rely on the linear relationship between PV performance parameters and meteorological parameters [59, 60] are often used to extrapolate field measurements to pre-defined conditions.

One of the most popular regression models is PVUSA [61, 62]. The model requires selecting measurements at high G_I on the plane of array (POA) (equal or greater than 800 W/m^2), fitting measurements of P_A or P_{TU}, G_I, T_{am} and S_W to Eq. 2.1 and estimating the coefficients β_1, β_2, β_3 and β_4 via OLS. The model assumes that I_A, primarily depends on the G_I, and the V_A, primarily depends on the T_m; which in turn, depends on the G_I, T_{am} and S_W. The coefficients are estimated for every monthly block of data and then monthly power ratings at PVUSA test conditions (PTC) are calculated by prediction of Eq. 2.1 at G_I = 1000 W/m^2, T_{am} = 20 °C and S_W = 1 m/s.

$$P_{MPP} = G_I(\beta_1 + \beta_2 G_I + \beta_3 T_{am} + \beta_4 S_W) \tag{2.1}$$

The model is accurate for c-Si PV, but not for thin-film technologies.

A modified model for thin-films was proposed in [63], which uses Eq. 2.2 and adds a constant loss factor, with β_1 through β_5 the regression coefficients. Measurements at G_I equal or greater than 50 W/m^2 could be used, widening the predictive application of the model. An indirect advantage of this modified model is that less data is filtered out, resulting in more accurate realization of the temporal characteristics of the PV system inside a larger operating window.

$$Y_A = G_I(\beta_1 + \beta_2 G_I + \beta_3 T_{am} + \beta_4 S_W) - \beta_5 \tag{2.2}$$

A method devised by NREL to estimate the degradation rate [64] with PVUSA, restricted data points to $G_I > 800$ W/m^2 and $P_A > 0.75 P_{nom}$ in order to eliminate data points where the PV array was shaded or covered. Then, the P_{PTC} ratings were calculated on a monthly

basis. The results of this work have shown that an amorphous silicon (a-Si) system suffered a rapid degradation of 7.25 % through its first year of operation. In the subsequent 5 years, the system has shown a degradation of 1.73 %/y.

Normalized and scaled metrics

Normalized [65] and scaled [51, 66] metrics are used for direct comparison between different PV technologies, PV plant capacities and geographical locations [67]. The most popular metric used is *PR*, which is defined in Eq. 2.3 as the ratio of the Y_A, or the final yield, Y_f, of the PV array/plant, and the reference yield, Y_r [36].

$$PR = \frac{Y_A}{Y_r} \tag{2.3}$$

where the Y_A is the sum of the array energy, $\sum E_A$, divided by the P_{nom} of the deployed PV array. The Y_r is the sum of the global irradiation, H_I, on the POA, divided by the global irradiance at STC, G_{STC}. This represents an equivalent number of hours at G_{STC}. Y_r can therefore be used to define the solar irradiance resource for the PV system. It is a function of the location, orientation of the PV array and weather variability.

PR ratings are defined as aggregates on a monthly or annual basis and are typically reported as such. Values calculated for shorter intervals, such as weekly or daily, may be useful for identifying component failures and measurement outliers. On the other hand, annual *PR* ratings can very easily be used to indicate a permanent loss in performance, i.e. degradation. The *PR* can also indicate losses due to soiling and seasonal variations when calculated on a monthly basis. This is one of the main advantages of *PR* and other normalized and scaled metrics over extrapolated metrics. Additionally, by normalizing with respect to irradiance, the overall effect of losses on the rated output due to 1) inverter inefficiency, 2) wiring mismatches and other losses from DC/AC conversion, 3) module temperature, 4) reflection losses. 5) soiling, 6) system down time and component failures can be captured in the metric.

With respect to degradation rates, a comparison of the published results obtained using PVUSA and *PR*, has shown similar results between them, for various PV technologies [68], but some studies have shown substantial differences between using the PVUSA and *PR* [65, 69]. Differences have also been observed when using temperature correction and various time units on the *PR*. It has been shown that the application of temperature correction resulted in higher estimated R_D, in comparison to using uncorrected measurements[65, 70]. This can be somewhat correlated to the instability of the temperature coefficients in the field which were reported to be non-linear [71, 72, 73]. Lastly, it was shown that using a smaller time grid with these metrics resulted in increasing the variability of the estimated R_D [65].

2.2.2 Outlier detection and treatment

An outlier is defined as an observation that is some measure of distance away from other data points and may indicate bad data, measurement error or faults in the system. When data is used to model a PV device, the model can only be as accurate as the data used during training. For this reason, an outlier can have a significant impact, depending on the analysis methodology. One way to reduce the number of outliers in PV measurement data is to use calibrated and accurate sensors in the field and adhere to a maintenance plan. Even then, outliers will surely be presented in the measurement data sets, due to the unpredictable state of the prevailing meteorological conditions, random interactions with nature and random system states.

A large number of publications can be found in the literature, that deal with detecting outliers in the data [74, 75, 76] and faults in the system [55, 77]. Outliers in the measurement data, especially due to shading [54], have been successfully identified using the I_{SC} and the G_I as a method to reduce uncertainty in R_D estimations [78].

Even though outliers and faults can be detected and the results used for planning maintenance actions and increasing the reliability of the system, this information does not currently suggest best practices in dealing with outliers in the measurement data. Therefore, it is common practice to filter field measurements to select favourable meteorological conditions [79] and average them to reduce noise and the effect of outliers. A pitfall in this approach is that excessive filtering of PV measurement data can lead to unrealistic expectations and introduce bias [80]. Also, it reduces sample size and the statistical significance of the results [81].

A more recent study discussed how data filtering affected the estimated degradation rate of a grid-connected PV system at NREL [65]. The study used the degradation rate estimated by measuring the P_{MPP} indoor at STC, prior to field deployment and after six years of exposure and compared it with different filtering criteria. It was found that the temperature corrected P_{MPP}/G_I and PR metrics, combined with a stability filter (i.e. irradiance and temperature rate of change), outlier filter and linear regression, resulted in degradation rates in line with the indoor measured degradation rate at STC. The non-temperature-corrected PR showed negative bias on the resulting degradation rate. Using the PVUSA rating the resulting degradation rates were found to be much different that the indoor degradation rate, independent of whether filtering was applied or not.

In other fields of engineering, robust statistical techniques [82] are used when the measurement data contains significant outliers. Robust techniques [83] were designed to mitigate their effect [84]. In addition, the detection of outliers is presented in a very different manner than most of the literature in the field of PV, apart from some excellent contributions [85, 86, 87] where statistical methods based on quantitative measures derived using the population sample are proposed.

Purely data-driven approaches such as principal component analysis (PCA) and its modifications are often taken into consideration [88]. PCA and the subspace methods classify

the observed data into "normal" and "abnormal" subspaces and have proven themselves efficient in anomaly detection applications [89, 90, 88]. Analyses such as z-scores on univariate data, Gaussian mixture models optimized using expectation-maximization, projections of the data onto lower dimensions (e.g. PCA and robust principal component analysis (RPCA) [91]), clustering and others can be applied if desired [92].

2.2.3 Missing data imputation

Measurement outages cause missing measurements and incomplete data sets. The outages can be due to data logger faults, measurement noise, data transmission errors, connection errors and data storage faults. Missing data is also caused by discarding invalid measurements and outliers due to inverter clipping, misconfigured sensors, sensor drifts and invalid calibration, array or irradiance sensor shading and soiling and PV system downtimes. Additionally, outliers can be determined from abrupt changes and high-frequency content in the measurements and also from highly volatile measurement data (higher than the response delay of the sensors.)

A power industry best practice to impute bad/missing smart meter data is presented in [93]. Intervals shorter than two hours were typically imputed by applying linear interpolation (LI) to the surrounding data. For periods longer than two hours, the typical approach was to construct daily load profiles based on previously validated historical data of similar days.

Similar approaches are common in the PV sector, where the subject of missing data imputation is very new, with a very small number of focused contributions. The most common approach in the PV sector to handle missing data entries is to ignore them completely. These methods include list-wise deletion and pairwise deletion. Even though these methods were the easiest to deal with empirically, their application introduced bias and assumptions that rarely hold true, such as lossless relationships between meteorological parameters and P_{MPP} and biased estimates in statistical analyses such as linear regression [94].

Some initial work has been done to impute time series using PV system models such as PVUSA [57] and regression models [95] in order to estimate the R_D. Another study reported the analysis of the performance of PV arrays using imperfect or incomplete input data[96]. The study proposed interpolation of missing meteorological data by calculating the nominal operating cell temperature, T_{NOCT}, using available measurements, and identifying erroneous PV measurement data by plotting. Most recently, empirical electrical models have been used [97] to back-fill time series of measurements. The basic idea between the two methods is the same: utilize as many covariates as possible from meteorology and field performance, in order to build a PV model that would accurately interpolate the series.

On the other hand, methods for handling missing data points is a well-established area in statistics [98, 99]. Full data sets can be generated by filling in the missing data periods with imputed data [100]. The imputed data periods have a continuous profile with respect to the adjacent available measurements, which is a highly desirable feature for time series

analyses. Common data imputation methods are categorized as single imputation (SI), multiple imputation (MI), multiple overimputation (MO) and maximum likelihood estimation (MLE) [101, 98, 102]. Modern MI approaches have been developed to impute values multiple times [98], i.e. multiple imputations, which can provide a measure of the uncertainty of the missing data point. In this way, a confidence interval could be constructed.

SI methods, such as replacing the missing values by the mean of the available values, by using linear regression to interpolate missing data and filling in gaps with the last observation carried forward (LOCF), are simple to implement, but can lead to biased estimates of certain parameters in statistical modeling. Compared to SI methods, MI and MLE methods have better statistical properties, but require much more computational resources.

2.3 Time series analysis

Statistical methods for estimating the trend of the performance metric over time have been shown to greatly affect the estimation of R_D [103]. The goal of the statistical analysis is to extract the trend of PV performance time series and translate the rate of change of the trend to the annual R_D. Model-based methods such as linear regression (LR), classical seasonal decomposition (CSD), Holt-Winters exponential smoothing (HW) and autoregressive integrated moving average (ARIMA) require the specification of a stochastic time series model whereas non-parametric methods, such as locally weighted smoothing (LOcal regrESSion) (LOESS) and Theil-Sen (TS) do not require the specification of a model and are popular because of their ease of use and robustness.

2.3.1 Linear regression

The most commonly used method in the literature is LR. It is used to fit Eq. 2.4 to the PV performance metric time series:

$$\hat{y} = \beta_1 t + \beta_2 + \epsilon \tag{2.4}$$

where $\hat{y}$ represents the fitted values, β_1 is the slope of the trend and β_2 is the intercept. The LR algorithm tries to fit Eq. 2.4 by minimizing the sum of squared residuals, most commonly by using OLS. It is very sensitive to outliers and seasonal variation and can thus exhibit a high uncertainty.

Other methods, more advanced than LR have been proposed in the literature [104, 105], to extract the underlying trend from PV performance time series and overcome the limitations of the LR method.

2.3.2 Theil-Sen estimator

An approach for estimating the R_{D_E} has been proposed by SunPower® [106, 31]. The so-called year-over-year (Y-o-Y) approach is an alternative to OLS and is known in the statis-

tical field as the Theil-Sen estimator [107, 108].

The TS is a robust estimation technique that chooses the median slope among all lines passing through the data points. This estimator can be computed efficiently, and is insensitive to outliers [109]. The linear slopes are calculated as follows:

$$d_k = \frac{X_j - X_i}{j - i} \tag{2.5}$$

for $(1 \leq i < j \leq n)$, where d is the slope, X denotes the variable, n is the number of points and i and j are indices. TS slope is then calculated as the median from all slopes: $b = \text{Median}(d_k)$. A confidence interval for the slope estimate is also easily determined as the interval containing the middle 95 % of the slopes of all lines.

A seasonal TS slope estimator can also be defined, to estimate slopes for each season (months or days.) It is calculated as follows:

$$d_{ijk} = \frac{X_{ij} - x_{ik}}{j - k} \tag{2.6}$$

for each (x_{ij}, x_{ik}) pair, $i = 1, 2, \ldots, m$, where $1 \leq k < j \leq n_i$ and n_i is the number of known values in the ith season. The seasonal slope is the median of all values of d_{ijk}

2.3.3 Classical seasonal decomposition

Generally, CSD is regarded as a simple method of seasonal adjustment [110] as the decomposition is performed with minimal effort and computational needs. This technique also forms the basis for most of the modern decomposition methods [111]. CSD and other statistical methods have been used in the past to model grid-connected PV power production [112] and to determine degradation rates of PV modules [104, 70]. Despite that and due to the fact that the particular technique fits a predefined model, it doesn't take into account the particular characteristics of each time series and therefore, cannot optimally model each different PV system technology. Additionally, since the most basic assumption of stochastic models is that model residuals are uncorrelated, in most cases it cannot provide statistical inference, proven by the presence of autocorrelation in the model residuals.

An additive model, as in Eq. 2.7, or a multiplicative model, as in Eq. 2.8, can be specified, depending on the nature of the seasonal component,

$$\hat{y} = T_t + S_t + e_t \tag{2.7}$$

$$\hat{y} = T_t * S_t * e_t \tag{2.8}$$

where $\hat{y}$ are the fitted values, T_t represents the trend, S_t represents the seasonal component and e_t the residual component.

2.3.4 Autoregressive integrated moving average

The most advanced model-based method reported in the literature was multiplicative ARIMA [113, 114]. The ARIMA method is more flexible than classical decomposition methods since it can effectively deal with seasonal variations, random errors, outliers and level shifts through the specification of a model which removes all autocorrelation in the model residuals. The general model for multiplicative ARIMA is given in Eq. 2.9 and is abbreviated as ARIMA(p,d,q)(P,D,Q), where p is the auto-regressive (AR) order, d is the differencing order, q is the moving average (MA) order, P is the seasonal AR order, D is the seasonal differencing order and Q is the seasonal MA order.

$$\Phi(T)\Phi_S(T^S)\nabla^d\nabla_S^D y_t = \theta(T)\theta_S(T^S)e_t \tag{2.9}$$

In order to fit the optimal ARIMA model, the time series must first be checked for stationarity and then transformed using differencing to achieve stationarity. The lags p, q, P, Q of the model are determined by inspecting the autocorrelation function (ACF) and partial autocorrelation function (PACF). The model selection procedure can yield multiple models that fit the data well. The optimum model is the one with the lowest order (i.e. parsimonious), with the lowest mean squared error (MSE) and the lowest Akaike information criterion (AIC), corrected Akaike information criterion (AICc) or Bayesian information criterion (BIC).

ARIMA was used to study the sensitivity of the methodologies to outliers and data shifts [104]. Shifts in the measurement data were fixed by selecting a range of corrective scaling factors and minimizing the residual sum of squares (RSS) of the errors with respect to the various scale factors. The ARIMA(1,0,0)(0,1,1) model was pre-selected and the results have shown that it performed very well with respect to outliers and could be used to calculate similar R_D to the simple LR method with as few as two years of field measurement data in a semi-arid climate.

2.4 Non-linearity

The most common way in the literature for estimating degradation was by assuming a linear trend that quantified the long-term performance. Field tests have shown that the assumption of linearity could not be held in the beginning of outdoor exposure, especially for thin-film technology.

A recent study [9] has shown that some degradation modes such as discolouration can lead to a linear degradation, while other degradation modes can lead to distinctly non-linear degradation which appears as large drops in the time series (e.g., hot spots, solder bond failures, corrosion). Non-linearities were typically more easily observed in accelerated tests than in he field, due to the high stress levels. In addition, as detailed in Sec. 2.1, modules in the field can display a variety of degradation modes and this leads to difficulty in correlating

physical degradation to a deterministic trend in the time series.

Specifically for thin-film techonologies, the initial performance of a-Si technologies features an initial rapid decline which stabilizes after several months [115, 57].

The degradation rate of non-linearly degradating PV array has not been explored in the literature.

2.5 Uncertainty

Each methodology for calculating the degradation rate of PV systems in the field carries its own uncertainty. For example, the LR method on PR carries the uncertainty of the regression and the calculation of PR, which in turn carries the uncertainties of the V_A, I_A and G_I sensors used.

It was reported that correction of outages and filtering measurements in order to exclude low-quality data reduced the uncertainty of the degradation rate due to reducing the variance in the data [65, 116]. On the other hand, the short observation time, the presence of outliers in the measurements and the data shifts related with hardware changes, increased the uncertainty of the calculation [104]. A minimum testing period of 3 to 5 years was found to be necessary in order to obtain accurate R_D estimates from field measurements, due to seasonal variations and higher initial degradation [117]. More specifically, the uncertainty of the statistical method used to calculate the R_D was reduced with increased observation time, as more sample data was recorded and therefore random variations and seasonality had a smaller impact on the underlying trend.

In addition, the methodology used can affect uncertainty and therefore the width and shape of the R_D distribution. A higher uncertainty was related to higher variance and thus resulted in a broader distribution. Granata, Boyson, Kratochvil, and Quintana [118] found that the estimated R_D were within the experimental uncertainty, which meant that they were difficult to infer, given the specific measuring equipment and analysis method. The irradiance measurement carried the largest contribution to uncertainty. Translated into PR values, the uncertainty could reach up to 4.5 % [119]. This further proves the need for employing a methodology which minimizes the uncertainty, especially given that a linear R_D less than 0.67 %/y is required to satisfy long-term warranties [120].

2.6 Degradation rate estimation methods

Table 2.1 summarizes the most commonly employed analytical techniques for estimating R_D using field measurement data, as reported in the literature. From this table it can be seen that there is a wide array of methods used to estimate R_D.

Table 2.1: Most common R_D estimation methods reported in the literature.

Metric	Filtering	Analysis	Ref.
Monthly W/W_p, from STC corrected P_A	AM1.5 and noon	PE between sequential Junes	[69]
Monthly PR	none	LR	[69]
Monthly PR	$G_I > 800\,\mathrm{W/m^2}$, outages	LR	[69]
Monthly P_{PTC}, using P_A	none	LR	[69]
Monthly P_{PTC}, using P_{MPP}	$G_I > 800\,\mathrm{W/m^2}$	LR	[27, 104]
Monthly P_{PTC}	$G_I > 800\,\mathrm{W/m^2}$	LR on top of CSD	[104]
Monthly P_{PTC}	$G_I > 800\,\mathrm{W/m^2}$	CSD on top of ARIMA(1,0,0)(0,1,1)	[104]
Monthly PR	outages	LR on top of optimal sARIMA	[103]
Monthly PR	outages	HW	[103]
P_{MPP} from module IV	none	PE	[58]
P_{MPP} from indoor IV	none	PE	[58]
P_{MPP} from module IV	$T_m = T_{NOCT}$ and $G_I > 800\,\mathrm{W/m^2}$	PE	[11]
Annual AC PR	none	PE	[121]
P_{MPP} from indoor IV	none	PE	[122, 123]
Weekly means of P_{MPP}, I_{SC}, V_{OC}, FF, I_{MPP} and V_{MPP} from module IV	extrapolation to $1000\,\mathrm{W/m^2}$ and $T_m = 45\,^{\circ}\mathrm{C}$ and $800 < G_I < 1100\,W\,m^{-2}$	LR	[124]
Daily Y_f	none	LR	[24]
Modified weekly and monthly P_{PTC}	extrapolation to STC and $G_I > 500\,\mathrm{W/m^2}$	LR	[125]
Performance Index = (Measured / Expected output)	$400\,\mathrm{W/m^2} < G_I < 2000\,\mathrm{W/m^2}$, bad data	TS	[106, 126]

The distribution of the 196 R_D values from all sources cited in this chapter can be seen in the histogram of Fig. 2.1, where a positive R_D indicates performance loss. Since the distribution of the published R_D figures is not exactly normal, due to limited sample size (196) relative to all deployed PV plants in the world, the mean and median values were bootstrapped 1000 times to reduce bias. The mean R_D of all technologies was estimated as (1.127 ± 0.140) %/y and the median as (0.989 ± 0.089) %/y at the 95 % confidence level. Similarly, for individual technologies, the mean R_D for monocrystalline silicon (mono-Si) was (0.969 ± 0.254) %/y, for polycrystalline silicon (poly-Si) was (0.818 ± 0.176) %/y, for a-Si was

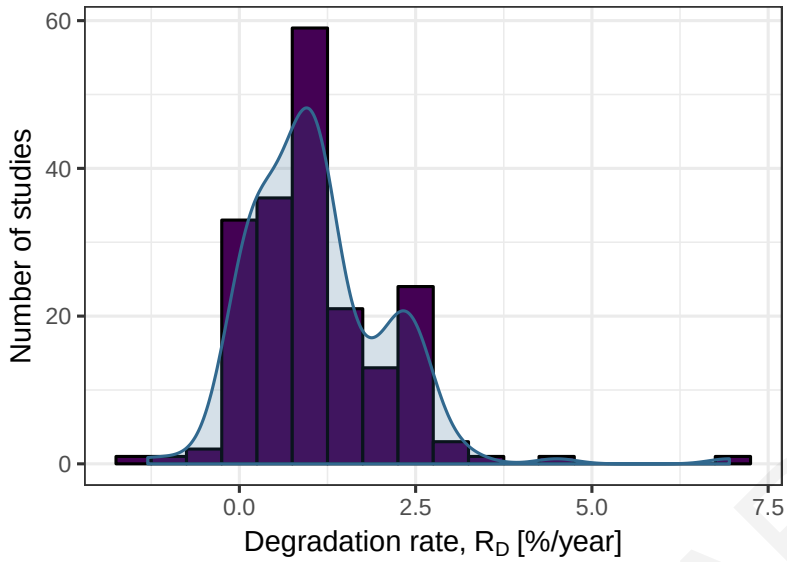


Figure 2.1: Distribution of degradation rates reported in the literature.

(1.413 ± 0.424) %/y, for copper indium gallium (di)selenide (CIGS) was (1.800 ± 0.478) %/y, for cadmium telluride (CdTe) was (1.657 ± 0.596) %/y and for other thin-film technologies the mean R_D was found to be (2.244 ± 0.087) %/y. From the differences in mean degradation rates, it is evident that they are highly dependent on PV technology. Degradation rates estimated specifically with LR are shown in Fig. 2.2, categorized by module technology.

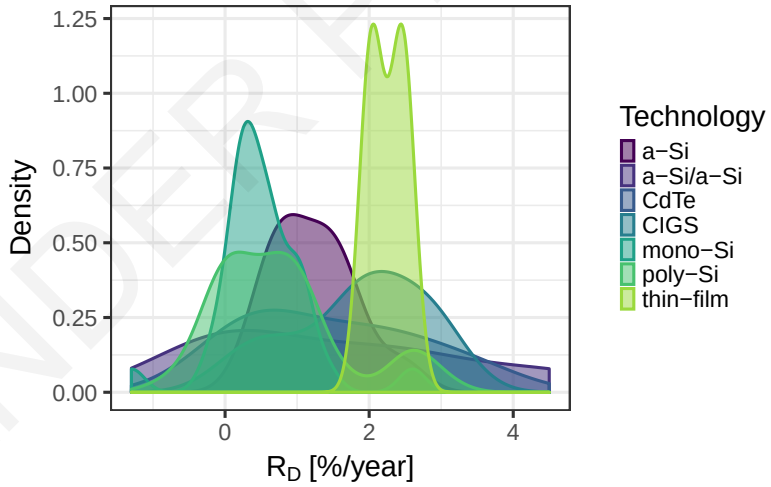


Figure 2.2: Distribution of degradation rates reported in the literature as estimated with LR.

The R_D results extracted from the literature were further categorized by statistical analysis method and are presented in Fig. 2.3, where the dashed horizontal line represents the median for all technologies. It can be seen that the IV method with PE resulted in the lowest degradation rates and low variation, except for mono-Si technologies. LR resulted in the highest variation and uncertainty, especially for a-Si-2J, CdTe and CIGS, and produced slightly lower median R_D. LOESS and ARIMA, albeit less frequently used, were shown to produce better results with low variation, for all technologies. Lastly, CSD produced the highest degradation rates for c-Si technologies.

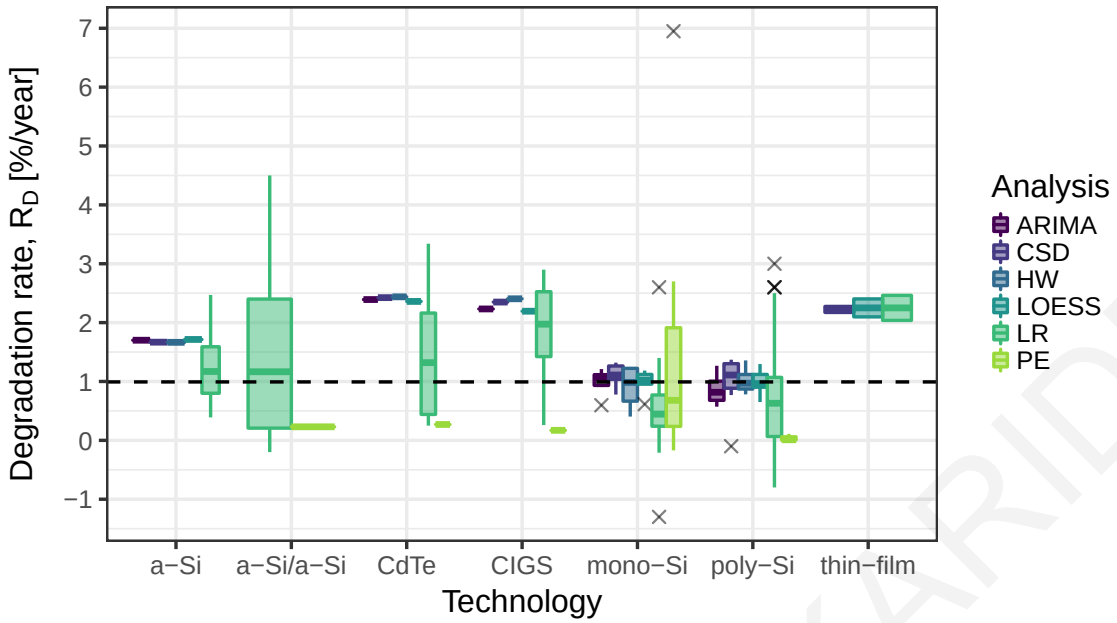


Figure 2.3: Degradation rates from the literature, categorized by technology and statistical analysis method.

2.7 Conclusions

In this chapter, it has been shown that the R_D did not only depend on PV technology, but it also depended on the analysis methodology. Many different methodologies for the estimation of R_D were found in the literature. The four major statistical analysis methods were: 1) LR, 2) CSD, 3) ARIMA, and 4) PE. The most commonly used performance metrics in conjunction with these methods were found to be: 1) electrical parameters from IV curves recorded in outdoor or indoor conditions and corrected to STC, 2) extrapolated metrics such as PVUSA, 3) normalized metrics such as PR and P_{MPP}/G_I, and 4) scaled metrics such as P_{MPP}/P_{nom}, and Y_A/P_{nom}.

The results of the reported studies have shown that the IV method with PE produced the lowest R_D. The LR method produced results with considerable variation and uncertainty. The CSD method produced the highest R_D for c-Si technologies but with lower uncertainty than LR, whereas the ARIMA and LOESS methods, albeit less popular, produced results with lower variation and uncertainty and good agreement between them for all technologies.

Finally, many deficiencies and gaps have been recognized. The gaps in the literature reveal a common pattern of constructing ad hoc analyses to solve statistical estimation problems.

Chapter 3

Experimental Setup

Work from this chapter has been published in [127]

3.1 Photovoltaic Technology test site

3.1.1 Systems and devices under test

At the outdoor test site of the PV Technology Laboratory of the University of Cyprus (UCY), eleven grid-connected PV systems of different technologies and approximately 1 kW capacity each, were installed and commissioned in June 2006. The performance of each PV system as well as the prevailing meteorological conditions were recorded since commissioning. Measured Meteorological quantities include the G_I, wind direction, a_W, S_W, as well as T_{am}, and T_m. The primary electrical quantities measured include V_A, I_A and P_{TU}. The characteristics and the P_{nom} for each PV array under study are listed in Table 3.1. Manufacturer specifications written on the datasheet for each technology are listed in Table 3.2.

Table 3.1: Characteristics of the PV arrays under study.

System	Cell Technology	No. of Modules	Nominal Power [kW_p]	γ_P [%/°C]
ucy04	mono-Si	7	1.540	-0.43
ucy05	mono-Si (HIT)	5	1.025	-0.30
ucy06	mono-Si	6	1.020	-0.37
ucy07	mono-Si (back-contact)	5	1.000	-0.38
ucy08	poly-Si (EFG)	4	1.000	-0.47
ucy09	mono-Si	6	1.110	-0.50
ucy10	poly-Si	6	0.990	-0.47
ucy11	poly-Si (MAIN)	6	1.020	-0.47
ucy12	CIGS	12	0.9	-0.36
ucy13	CdTe	18	1.08	-0.25
ucy14	a-Si	10	1	-0.20

Table 3.2: Datasheet specifications of the PV modules under study.

System	Efficiency [%]	V_{OC} [V]	I_{SC} [A]	I_{MPP} [A]	V_{MPP} [V]	P_{nom} [W_p]	Area [m^2]
ucy04	13.4	36.5	8.25	7.62	28.9	220	1.64
ucy05	16.4	50.3	5.54	5.05	40.7	205	1.25
ucy06	12.9	44.0	5.10	4.75	35.8	170	1.32
ucy07	16.1	47.8	5.40	5.00	40.0	200	1.24
ucy08	11.7	70.9	4.91	4.55	57.1	250	2.14
ucy09	14.8	44.8	5.50	5.10	36.5	185	1.25
ucy10	12.7	43.9	5.10	4.60	35.5	165	1.30
ucy11	13	44.0	5.25	4.71	36	170	1.31
ucy12	10.3	45.5	2.50	2.22	36	75	0.73
ucy13	8.3	90.0	1.14	0.94	64	60	0.72
ucy14	6.4	141.0	1.17	0.93	108	100	1.57

The sensors included a Kipp & Zonen CM21 secondary standard pyranometer located in the middle of the testing field, mounted on the POA, a PT100 back-of-module temperature sensor for every PV array, wind speed and wind direction sensors, dc (array) voltage and current transducers and an ac energy meter, as listed in Table 3.3.

Table 3.3: Sensors in use at the PV Technology test site.

Apparatus	Manufacturer & Model	Standard Error
Data acquisition	Delphin TopMessage	±0.01%
Pyranometer	Kipp & Zonen CM21	±2%
Ambient temperature	Theodor Friedrich 2030	$\pm 0.1 + (0.005 \times T)/3$
Module temperature	Heraus PT100	$\pm 0.3 + 0.005 \times T$
Wind speed	Theodor Friedrich	0.3 m/s at $v < 10$ m/s
Wind direction	Theodor Friedrich	±2.5°
Voltage sensor	Bastizi potential divider	±0.015 % at 20 V to 1000 V
Current sensor	Bastizi shunt resistor	±0.75 % at 2 A to 10 A
AC energy meter	NRZ AAD1D5F	±1 %

Fig. 3.1 shows a closeup of the location of the sensors at the testing site of the PV Technology laboratory and Fig. 3.2a shows the typical location of a PT100 temperature sensor on the back of the module. The pyranometer was installed on the POA at the same inclination angle as the modules (27.5 ± 1.0)°. Fig. 3.2b shows the sensor board which was used to measure the dc current and voltage of the PV array. This board was installed in the electrical panel next to the inverter. Measurements were acquired by a Delphin TopMessage data logger, shown in Fig. 3.3. Each data logger channel was sampled at 50 Hz and the samples were averaged and stored every 1 s, 1 min and 15 min, for effective archival rates of 1 Hz, 16.67 mHz and 1.11 mHz respectively. The data logger pushed these average measurement values to the database server which stored them indefinitely. This helped facilitate the creation of a high resolution set of measurement data which captured effectively the variance

Figure 3.1: Closeup of sensors in the field.

(a) Back of module temperature sensor

(b) Sensor board used to measure I_A and V_A

Figure 3.2: Sensors interfacing with the PV modules.

of each measurand.

The standard error given in Table 3.3 represents the combined standard uncertainty of measurement, u, as the estimated standard deviation of the result. Using the data in Table 3.3, the expanded measurement uncertainty at 95 % confidence, U_{95}, was calculated using uncertainty propagation [128] as in Eq. 3.1:

$$U_{95} = k * \sqrt{\left(\frac{d}{dx_1}u(x_1)\right)^2 + \left(\frac{d}{dx_2}u(x_2)\right)^2 + \ldots + \left(\frac{d}{dx_n}u(x_n)\right)^2} \tag{3.1}$$

where $u(x_1), u(x_2), \ldots, u(x_n)$ were the individual standard uncertainty components and k =

Figure 3.3: Delphin TopMessage data logger at the outdoor test site.

2 was the coverage factor for 95 % confidence. The expanded uncertainty for the V_A was calculated at ±0.036 %, for the I_A at ±1.5 % and for the $P_{TU} = V_A * I_A$ this was calculated at ±1.6 %.

3.1.2 Sampling and archival rates

General guidelines

At a minimum, the sampling rate of measurement should satisfy the Nyquist–Shannon sampling criterion [129]. The Nyquist–Shannon sampling theorem states that all the information from a continuous-time signal of finite bandwidth could be captured in a discrete sequence of samples, given a sufficient sampling rate. Such a sampling rate is only considered sufficient when the fidelity of the original signal is preserved during reconstruction of the continuous-time signal from its discrete samples [130]. Thus, the Nyquist frequency is defined as half the sampling rate and is therefore the maximum frequency that can be detected for a given sampling rate. Since in this work the signals were sampled at 50 Hz and archived at 1 Hz, the maximum frequency that could be detected was 0.5 Hz.

Literature which investigates the important subject of sampling and optimal PV performance measurement is sparse [131, 132]. The most widely accepted guideline, IEC

61724:1998 [36], recommends that the sampling interval for parameters which vary directly with irradiance shall be 1 min or less. Given that solar irradiance is highly volatile, a 1 min sampling interval is very large and thus it can cause aliasing. Even though the intent of IEC 61724:1998 was not to capture transient-level detail, but rather to suggest a sufficient sampling rate for characterizing performance over an averaging interval [133, 134], aliasing caused by an insufficient sampling rate can in turn cause unexpected deviation from the true value when the measurand varies and thus introduce error in average or integrated quantities.

The most important signals in PV monitoring are those that represent power: G_I, P_A and P_{TU}. These signals are integrated over time to enable the calculation of energy, as each value stored at a certain archival rate represents the average over that interval. System KPIs such as net energy from array, E_A, net energy to utility grid, E_{TUN} and PR are then derived from these quantities. In this work, since the commissioning of the systems under study in June 2006, PV system and meteorological quantities were sampled by the data logger at 50 Hz, then averaged and archived at 1 Hz, 16.67 mHz and 1.11 mHz, as previously mentioned. This far exceeded the sampling guidelines of IEC 61724:1998, for the purpose of creating a high-resolution data set of PV performance which captured transient effects such as volatility due to the weather [135] and MPP tracking step response [136]. The need for this highly resolved data diminishes when the goal of the analysis is to extract global features, such as energy yield and degradation rates.

Simulated sampling rates

Since the highest archival rate was 1 Hz, a one month long selection of 1 Hz averaged data from the ucy05 system, as well as the prevailing weather conditions, was extracted from the database in order to investigate whether the data could sufficiently characterize PV performance. Realistically, due to the 50 Hz sampling rate, aliasing due to sampling was not expected. Nevertheless, different sampling rates were simulated (based on the 1 s data) in order to assess whether strict compliance with the sampling guidelines in IEC 61724:1998 would be sufficient. For the purpose of this analysis, 1 Hz data was decimated [137, 138] by a factor of 10, 60 and 900 to simulate sampling intervals of 10 s, 1 min and 15 min respectively.

Analysing data at various sampling rates could show whether noise was introduced at excessively high sampling rates and whether frequencies were lost at low sampling rates. A low-pass filter is required first in the chain to mitigate distortion due to aliasing by downsampling. Three types of infinite impulse response (IIR) filters are typically used as anti-aliasing filters: Butterworth, Chebyshev Type I and Chebyshev Type II filters. Butterworth filters feature a maximally flat frequency response in the passband but weak roll-off, whereas Chebyshev Type I filters feature ripple in the passband and steeper roll-off and Chebyshev Type II filters feature flat passband response, ripple in the stopband (frequency leakage) and steeper roll-off. The frequency response of these filters for a cut-off frequency of 0.1 Hz and filter orders of 2, 4 and 8 is shown in Fig. 3.4. Of the three filter types, the

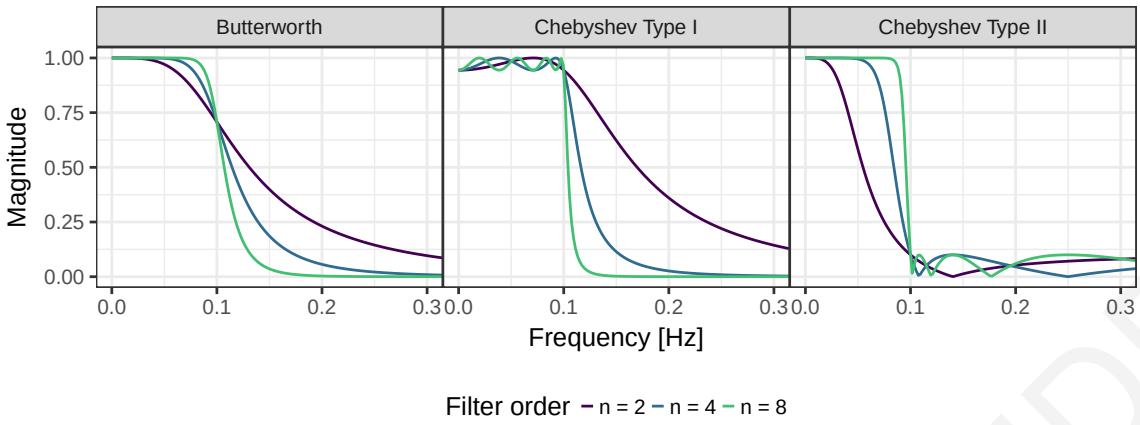


Figure 3.4: Frequency response of Butterworth, Chebyshev Type I and Type II filters.

Chebyshev Type I filter's features were the most desirable for this application. The filter was applied in both the forward and reverse directions to remove phase distortion. The gain response, $G_n(\omega)$, of the filter is given by Eq. 3.2:

$$G_n(\omega) = \frac{1}{\sqrt{1 + \frac{\epsilon^2 T_n^2 \omega}{\omega_0}}} \tag{3.2}$$

where T_n is a Chebyshev polynomial of nth order. In this case, $n = 8$ was chosen.

Data was decimated by first reducing high frequency signal components with a low-pass filter of cut-off frequency 0.1 Hz, 16.67 mHz and 1.11 mHz and then downsampling. A comparison of the 1 s data to the downsampled 10 s, 60 s and 15 min is shown in Fig. 3.5, Fig. 3.6 and Fig. 3.7 respectively. Through visual analysis, it can be seen that beyond 10 s sampling interval, some amount of information is lost.

Spectral analysis

To observe the frequency content of each signal, the discrete Fourier Transform (DFT) was computed using the Fast Fourier Transform (FFT). The DFT converts the signal from the time domain into the frequency domain and is given by Eq. 3.3:

$$X_k = \sum_{n=0}^{N-1} x_n e^{-i2\pi kn/N} \tag{3.3}$$

where:

- x_n is a sequence of uniformly spaced samples of signal $x(t)$,
- $k = 0, \ldots, N-1$,
- N is the number of samples.

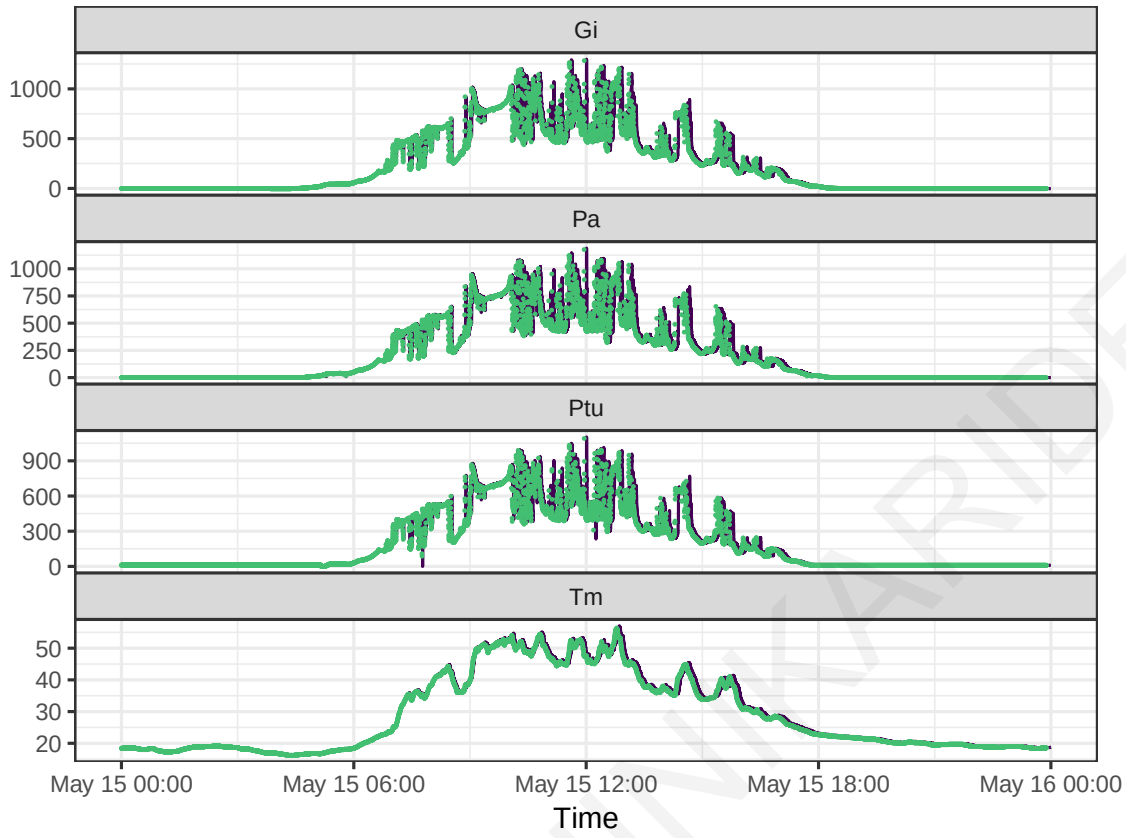


Figure 3.5: Comparison of 1 s archived data to 10 s downsampled signals.

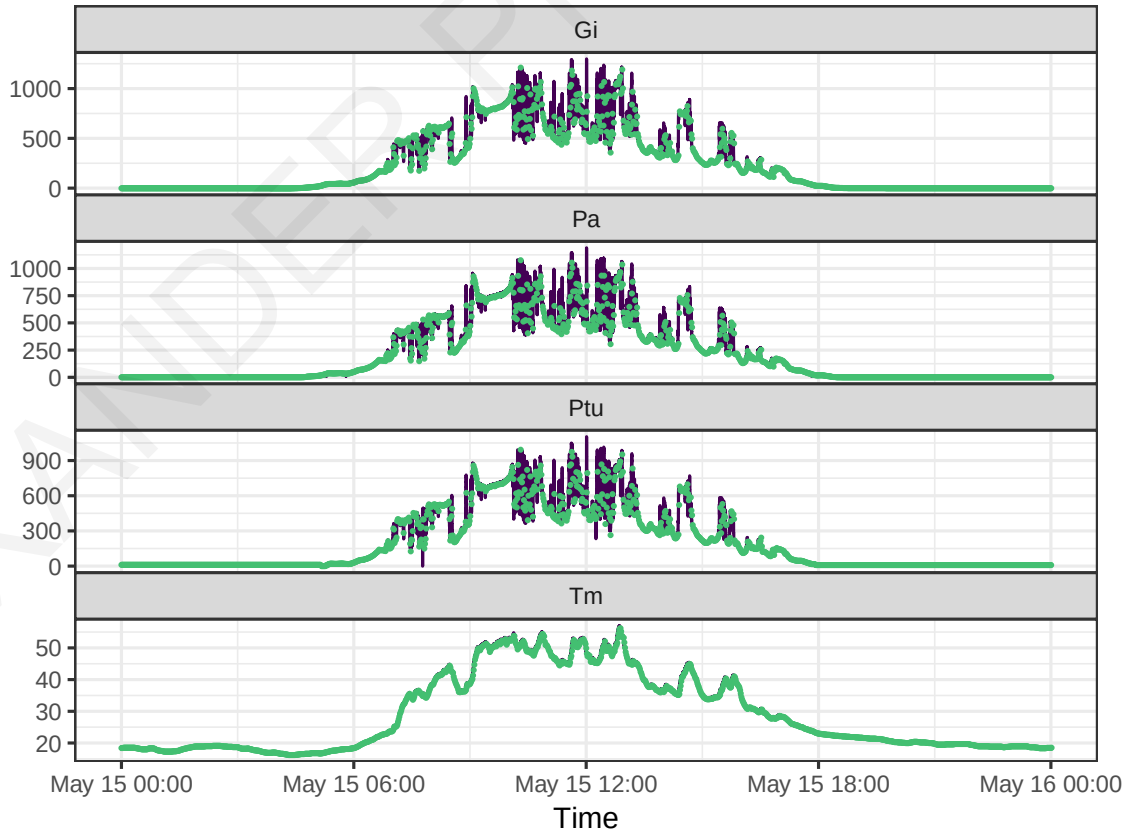


Figure 3.6: Comparison of 1 s archived data to 60 s downsampled signals.

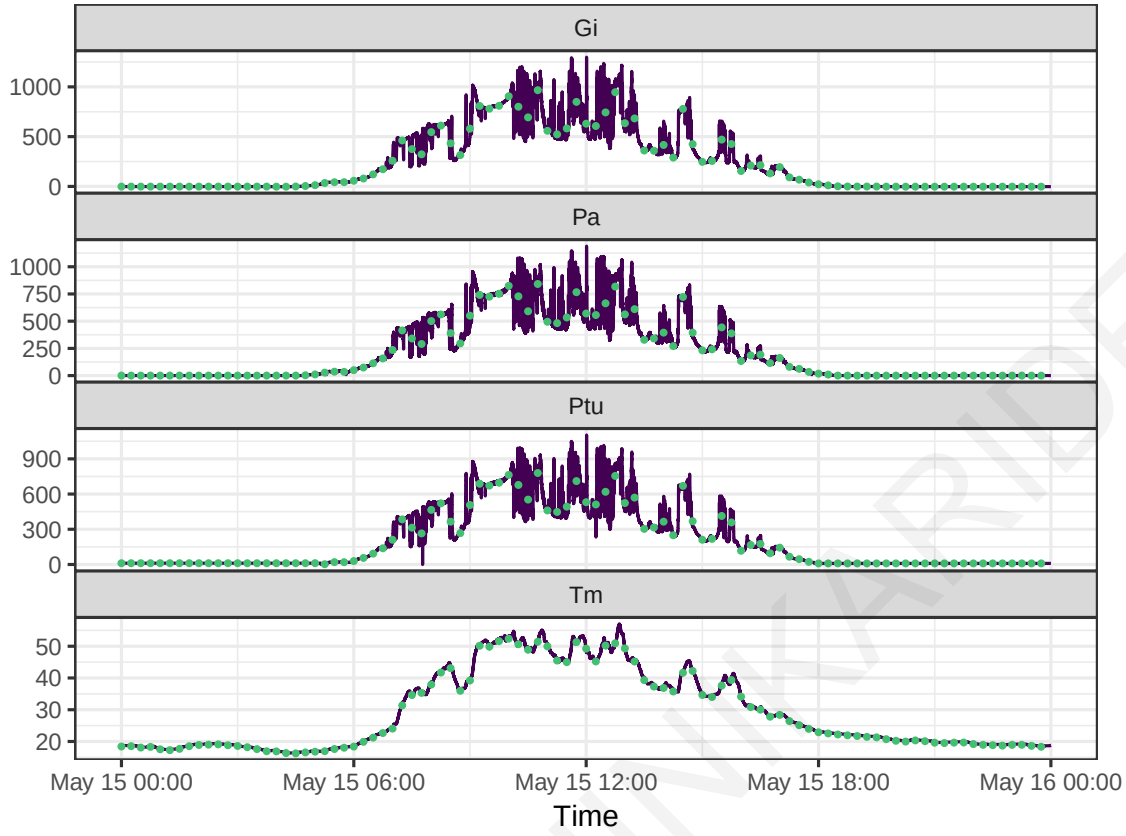


Figure 3.7: Comparison of 1 s archived data to 900 s downsampled signals for a single day.

The magnitude of the DFT was then computed, squared and plotted versus frequency to calculate the power spectral density (PSD) by Eq. 3.4:

$$S_{xx}(f) = \frac{\Delta t}{N}\left|\sum_{n=0}^{N-1} x_n e^{-i2\pi f n}\right|, \quad -1/2\Delta t < f \leq 1/2\Delta t \tag{3.4}$$

where Δt is the sampling interval. For a one-sided plot, the values at all frequencies except 0 and $1/2\Delta t$, are multiplied by 2 so that the total power is conserved.

Since the PSD shows at which frequencies variations are strong and at which frequencies they are weak, a linear trend and/or drift in the time domain will be converted to a high power component at extremely low frequency. Therefore, a linear trend and the mean were removed from the data prior to calculating the DFT and PSD by fitting a straight line with OLS and keeping only the residuals. From the PSD of 1 Hz P_A data, shown in Fig. 3.8, it can be seen that most of the information was contained at low frequencies and that there were short peaks and ripples at specific frequencies. Since the PSD was dominated by noise, Welch's method was used to reduce the effect of noise on the PSD plot. This method averages multiple spectra to compute the PSD by binning the time series at the fundamental period (i.e. 24 h) and then computing the PSD for each bin and averaging over all PSDs. The result can be seen in Fig. 3.9 which shows more clearly the locations of significant frequency components. The identified peaks on P_A at 0.2175 Hz, 0.328 Hz, 0.345 Hz, 0.4375 Hz and 0.454 Hz corresponded to periods of 4.60 s, 3.05 s, 2.90 s, 2.29 s and

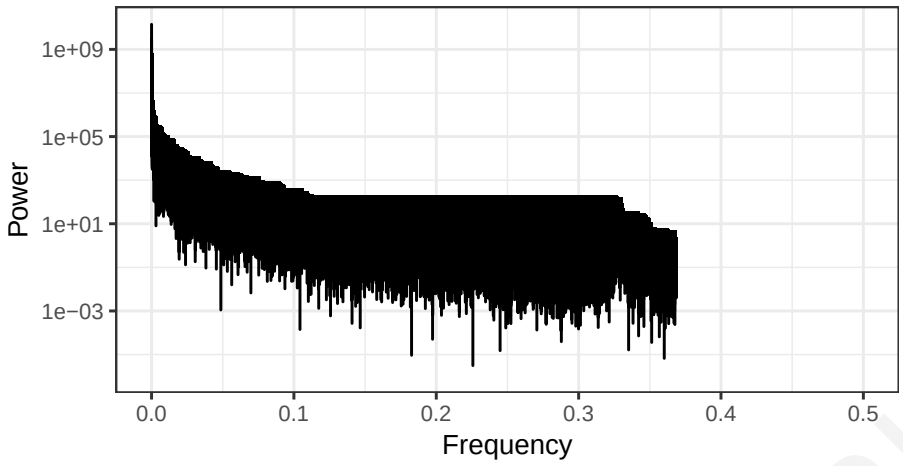


Figure 3.8: PSD plot of the detrended and demeaned 1 s P_A.

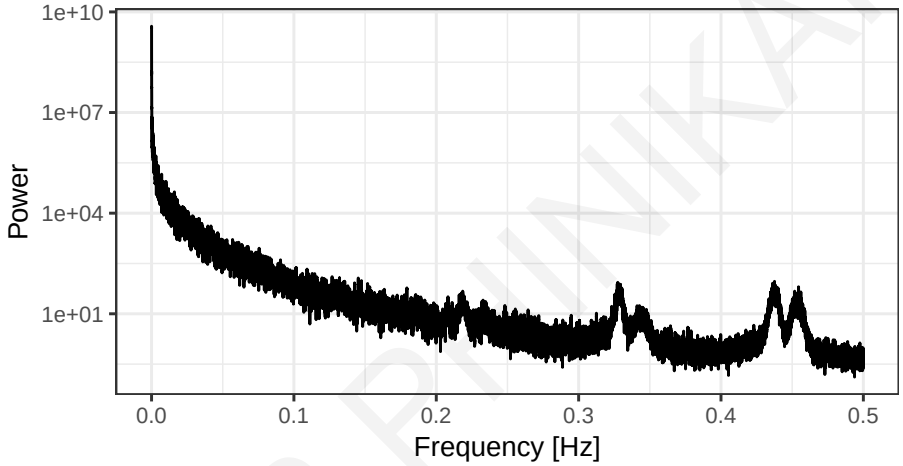
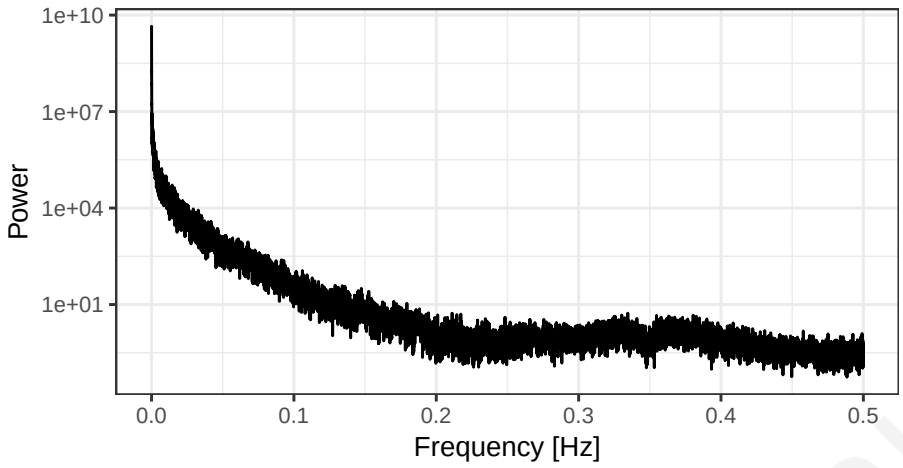


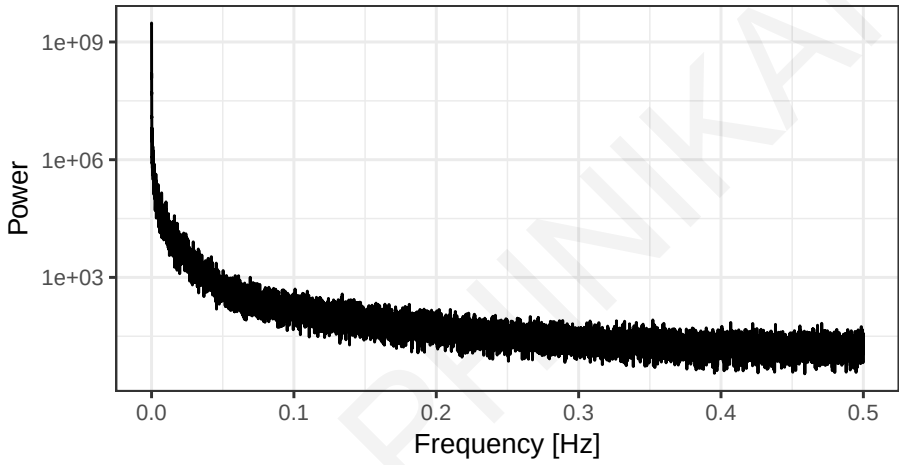
Figure 3.9: PSD plot of the detrended and demeaned 1 s P_A with Welch's method.

2.20 s respectively. These frequencies could only be detected on the P_A and not on G_I (see Fig 3.10a), P_{TU} (see Fig 3.10b) or T_m (see Fig 3.10c). An interesting feature of this analysis was the profile of these "high" frequency peaks, which were different for each sensor. As G_I was measured with a thermopile pyranometer which could not respond to fast irradiance changes because of its thermal mass, high frequency content were effectively filtered out by the sensor. Similarly, the PSD of T_m showed only noise due to inertia against temperature fluctuations. Regarding the frequency content of the P_{TU}, no peaks can be distinguished which meant that they were filtered by the inverter. Finally, the peaks shown in the PSD plot of P_A could be attributed to irradiance variations, since a PV panel/array would generally respond much faster than a thermopile pyranometer. To view the high frequency content in the time domain, a high-pass Chebyshev Type I filter with cut-off frequency $f = 0.1Hz$ was used with the 1 Hz P_A data to remove the low frequency content. A plot of the original data as well as the filtered high frequency content is shown in Fig. 3.11.

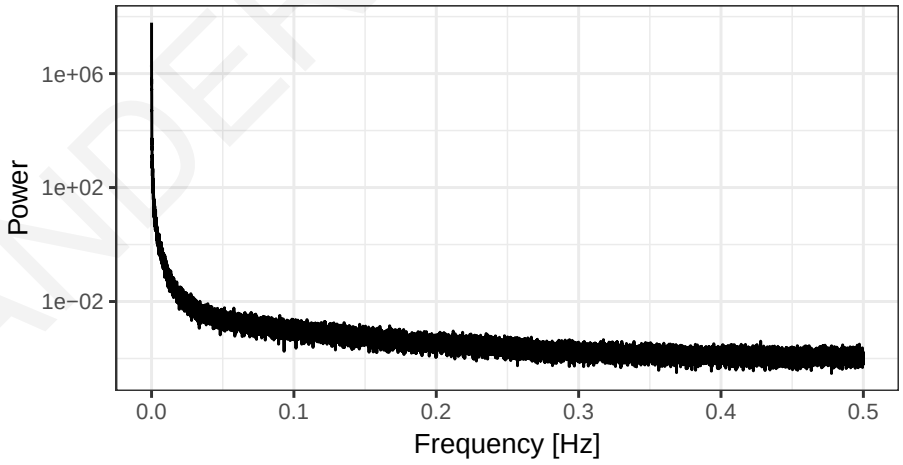
For completeness, PSD plots of P_A downsampled at 0.1 Hz, 16.67 mHz and 1.11 mHz were constructed. As expected, the plots have shown that high frequency content was missing. The only difference was on the 1.11 mHz (900 s interval) downsampled data which



(a) G_I



(b) P_{TU}



(c) T_m

Figure 3.10: PSD plots of the detrended and demeaned 1 s G_I, P_{TU} and T_m with Welch's method.

featured two low frequency peaks, at $f_0 = 1.157407 * 10^{-05} Hz$ and its harmonic at $f_1 = 2.314815 * 10^{-05} Hz$, corresponding to the diurnal and semi-diurnal cycles respectively. These are shown in Fig. 3.12.

In conclusion, it was shown that measurement data used in this analysis retained high

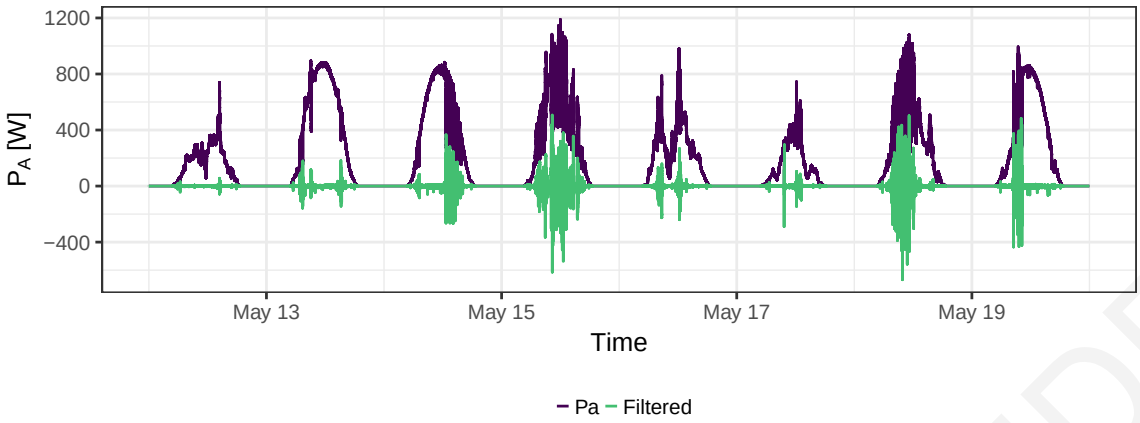


Figure 3.11: Original 1 Hz P_A data and its high frequency content.

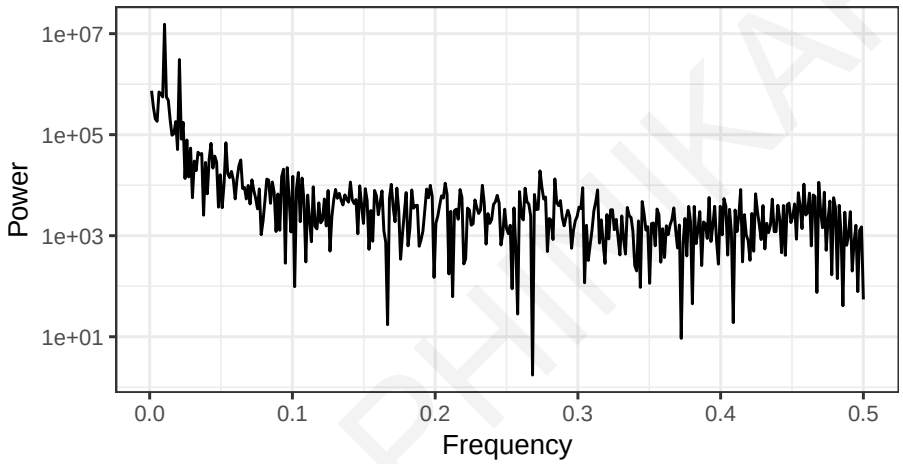


Figure 3.12: PSD plot of fifteen-minute downsampled P_A.

frequency variations that were lost when the sampling interval was increased to be in line with current practices and guidelines. Therefore, it was proven that the experimental setup used in this work could be used to analyze PV performance without being hindered by aliasing. The ability to record high frequency variations was dependent on the type of sensor used; sensors with thermal mass did not respond as fast as the PV panels/arrays to fast irradiance changes.

3.1.3 Quality assurance for measurements in the field

To ensure correct representation of the results, identify sensors which exhibit drifts and maintain measurement accuracy, regular calibration campaigns were performed for the resistive sensors in Fig. 3.1 which were used to measure the V_A and I_A as well as for the data logger and the Kipp & Zonen CM21 pyranometer.

Calibration was performed strictly on sunny days and under clear skies, to minimize measurement error from fast fluctuations of the I_A, due to clouds overhead. Calibration was performed at most every four to five weeks and all historical measurements were kept in a record with the absolute calibration and drift of each sensor.

Three people were involved in each calibration campaign. All three were required to carry synchronized time sources and record the measurement given by the apparatus they were responsible for at pre-defined intervals (usually every 10 s.) Multiple measurements were recorded for each type of variable and sensor which were then averaged in post-processing.

One person was responsible for measuring the V_A and I_A directly from the PV array cables while in operation. Another person was responsible for measuring the output of the array current and voltage sensors at the data logger. The third person was responsible for recording the value written into the database by the data logging software through a PC. Through this procedure, it was easy to detect the calibration of the sensors, pinpoint the cause of measurement errors, e.g. direct current (DC) cable losses, sensor cable problems, sensor element failures, data logger interface problems, misconfigured software calibration factors.

The historical calibration record of each sensor was examined for drifts after each calibration campaign and sets of calibration factors were calculated and used to correct the response of each sensor. Each set of calibration factors was applied on the appropriate data logger channels of V_A and I_A by adjusting the sensitivity. Therefore, this ensured that the measurements used throughout this work to calculate the fifteen-minute average P_A were accurate and the sensor variance was traceable.

Regarding module temperature, T_m, the measurements and the placement of the back-of-module temperature sensors were cross-checked with an accurate IR camera in order to ensure that the measured T_m was not skewed due to seasonal shading or vicinity to hot spots on the module.

Lastly, the CM21 pyranometer which was used throughout the evaluation period was calibrated every two years by the manufacturer or accredited calibration laboratories. Whenever the CM21 was dismounted for calibration, it was replaced by a calibrated secondary standard EKO MS-802 pyranometer. This ensured that measurements would continue to be recorded with the same standard.

3.2 Indoor test laboratory

The Pasan SunSim 3c solar simulator was used to measure the performance of each PV module at standard test conditions. IEC 60904-9:2007 [139] classifies sun simulators in three classes, A (the best), B and C depending on three parameters, summarized in Table 3.4. From

Table 3.4: Uncertainty of the solar simulator measurements.

Parameter	Class A	SunSim 3c
Non-uniformity of irradiance	≤ 2%	≤ 1.0%
Pulse instability (long term)	≤ 2%	≤ 1.0%
Spectral irradiance distribution	≤ ±25%	≤ ±12.5%

the above technical parameters, the SunSim 3c could be rated class AA – AA – AA (or A+ according to TÜV).

The expanded uncertainty of measurement including measured temperature uniformity, irradiance uniformity, spectral mismatch from AM1.5 using spectroradiometers and IR temperature sensors was calculated as ±3.5 % for the power at STC, P_{STC}.

3.2.1 Quality assurance for measurements in the lab

Large area pulsed solar simulator

The same standards of measurement quality and calibration were also implemented for the large area pulsed solar simulator (LAPSS), as it was the basis of the indoor experimental approach. Two new c-Si PV modules were designated as the reference modules, and were only used to verify calibration of the flasher. These two modules were sent initially once, and later twice a year, to the calibration centre of the Joint Research Centre (JRC) of the European Commission, which provided traceable calibration certificates with expected electrical characteristics at STC which were used to adjust the sensitivity of the flasher.

Periodic quality control

Prior to every PV module test at STC, one of the reference modules was mounted on the flasher and used to verify calibration. The results were recorded in the database and are graphed in Fig. 3.13, along with the 95 % confidence interval (CI).

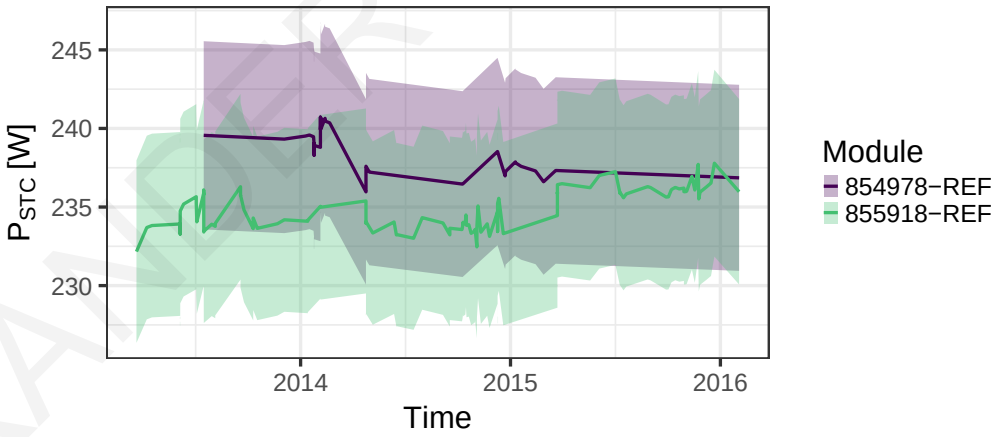


Figure 3.13: P_{STC} of the reference modules used to calibrate the indoor solar simulator.

Round-robin calibration

In addition to accredited calibration, two round-robin tests were performed throughout the evaluation period, to test the stability and calibration of the flasher with different PV technologies. The first round-robin module test was performed between the UCY, the Austrian Institute of Technology (AIT), the Institut für Photovoltaik, Universität Stuttgart (ipv) and the Zentrum für Sonnenenergie- und Wasserstoff-Forschung (ZSW). The scope of the

round-robin test was to accurately measure the PV modules' initial performance at STC and to perform full characterization by using common indoor testing procedures according to international standards. The end result was a calibration value of the testing equipment, adapted to each module type under study.

The first round-robin test was concerned with the accurate measurement of the IV curve of the PV modules at STC, the optical characterization to find defects and the comparison of the results between the test facilities. Testing was done on unexposed control modules: one Schott Solar ASI103 a-Si module, one TSMC TS150 CIGS module and one Enfoton 240QC poly-Si module.

For the accurate comparison of the thin-film modules' P_{STC}, the spectral correction factor had to be calculated and applied. This was done by in situ measurement of the module's I_{SC} at AM1.5 global spectrum and under high irradiation. The in-situ-measured I_{SC} was then extrapolated to STC and compared to the output of a calibrated c-Si reference cell, mounted in-plane. Due to the Earth's movement, AM1.5 global spectrum (AM1.5 G) is available outdoor only twice a year; February–April in Cyprus and April–May in Austria and Germany. These periods were calculated through simulation using the Michalsky sun position algorithm [140]. The equation for the air mass, AM is shown in Eq. 3.5:

$$\mathrm{AM} = \frac{1}{cos(\theta_z)} \tag{3.5}$$

where θ_z was the zenith angle at the PV testing site, calculated through the sun position algorithm.

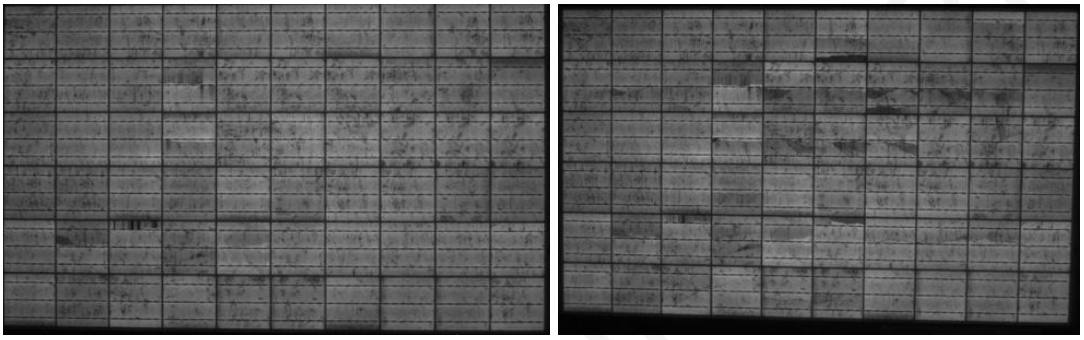
Testing at the UCY was performed in March 2014 and April 2015. Prior to testing at STC, the poly-Si module was exposed to 5 kW h/m^2 irradiation under load, in accordance to IEC 61215-1-2:2016 [141]. The corresponding standard for thin-film modules, IEC 61646:2007 [142], calls for light-soaking at successive intervals of 43 kW h/m^2 irradiation under load, until stabilization of P_{STC}, measured using the indoor solar simulator. In Cyprus, the P_{STC} was stable after 43 kW h/m^2 irradiation. While performing light-soaking on the modules, the modules' electrical characteristics, as well as meteorological parameters were being recorded. The desirable spectrum (AM1.5 G) was available outdoors and therefore the I_{SC} was measured as close and extrapolated to 1000 W/m^2, AM1.5 G and 25 °C. The extrapolated I_{SC} was then used to calculate the spectral mismatch factor (MMF) which was subsequently used to adjust the solar simulator sensitivity. The calculated spectral MMF varied by PV module technology:

- Schott Solar ASI103: 8.67 %
- TSMC TS150: 0.79 %
- Enfoton 240QC: 0.76 %

The spectral MMF from AM1.5 G for the TSMC and Enfoton modules were insignificant and were thus not applied in order to avoid introducing further uncertainty to flash tests.

On the contrary, the spectral MMF for the Schott Solar ASI103 module was significant and contributed greatly to correct measurement of the module rating at STC.

To further ensure that the results were consistent across all test laboratories, EL imaging was performed before shipping the modules and upon reception at their destination; the AIT or the ipv. Comparison of the EL images before shipping and upon reception would reveal damaged incurred by shipping and handling. It was expected that any damage on the PV cells would affect the measured P_{STC}. Fortunately, that was not the case during the first round-robin. But, during the second round robin, microcracks could be identified from the EL image, as shown in Fig. 3.14a and Fig. 3.14b.



(a) Before transport (b) After transport

Figure 3.14: Small microcracks from shipping and handling (a) before and (b) after reception back in Cyprus.

The results of the first round robin have shown very good agreement for the Schott Solar a-Si and TSMC CIGS module results between the UCY and AIT, at 99.85 % and 98.82 % respectively. The STC results for the Enfoton poly-Si module have shown up to 2.92 % difference between the two test labs, which was within the uncertainty of measurement. All three results were within the experimental uncertainty and validated the correct rating of the modules at STC. Similarly, the comparison of the results between UCY and ipv has shown maximum differences of up to 3.8 W.

One year later, the round-robin was repeated between the same research labs. The results have shown differences between 0.37 % to 0.95 % from the first round-robin. The differences were negligible and have therefore verified the stability of the experimental setup.

Chapter 4

Data Organization

4.1 PV performance measurements

Fifteen-minute average measurements of the V_A, I_A, T_m, and G_I were used to develop the work in this dissertation, to assess the R_{D_E} of each PV array in the field. The G_I was measured on the POA, using a calibrated Kipp & Zonen CM21 pyranometer. Using sampled V_A and I_A, fifteen-minute average P_A was calculated as $P_A = \frac{\sum_n V_A * I_A}{n}$, where $n = 15 * 60 * \frac{1}{T}$ and $T = 1/50\,\text{Hz} = 0.02\,\text{s}$ was the sampling interval. Measurements were extracted from the database for the period between 2006-06-01 and 2015-05-31, therefore the first nine full years of operation. Monthly PR time series, which were constructed using fifteen-minute average data in Eq.2.3, for the PV systems listed in Table 3.1, are shown in Fig. 4.1, from June 2006 to May 2015.

The PR time series presented in Fig. 4.1 have been produced using the fifteen-minute average measurements and manual monthly energy yield corrections (in kW h) based on the ad hoc back-filling procedure using similar days of measurements, which was described in Ch. 2. These historical corrections accounted for most of the energy lost due to system faults and outages and were being calculated since the commissioning of the systems in June 2006 for the purpose of energy yield comparisons between the technologies. They are plotted in Fig. 4.2 and are displayed here for completeness' sake, to visualize them alongside the monthly PR time series shown in Fig. 4.1. In Ch. 5, a methodology will be described that deals with this subject, in the context of the unsupervised degradation rate estimation.

The main requirements that enable the analysis presented in this work are good quality measurements of P_A and good quality measurements of G_I. Although PV reference cells could also be used as irradiance sources (to offset the cost of a quality pyranometer) [143] their measurements can be affected by the same degradation modes as conventional PV modules. The presence of a temporal drift in the output of the reference cell, due to degradation, would make the estimation of the R_{D_E} impossible.

Lastly, metadata such as the GPS coordinates of the site of installation of the PV system, as well as the timezone of the location and the angle of inclination of the PV array were

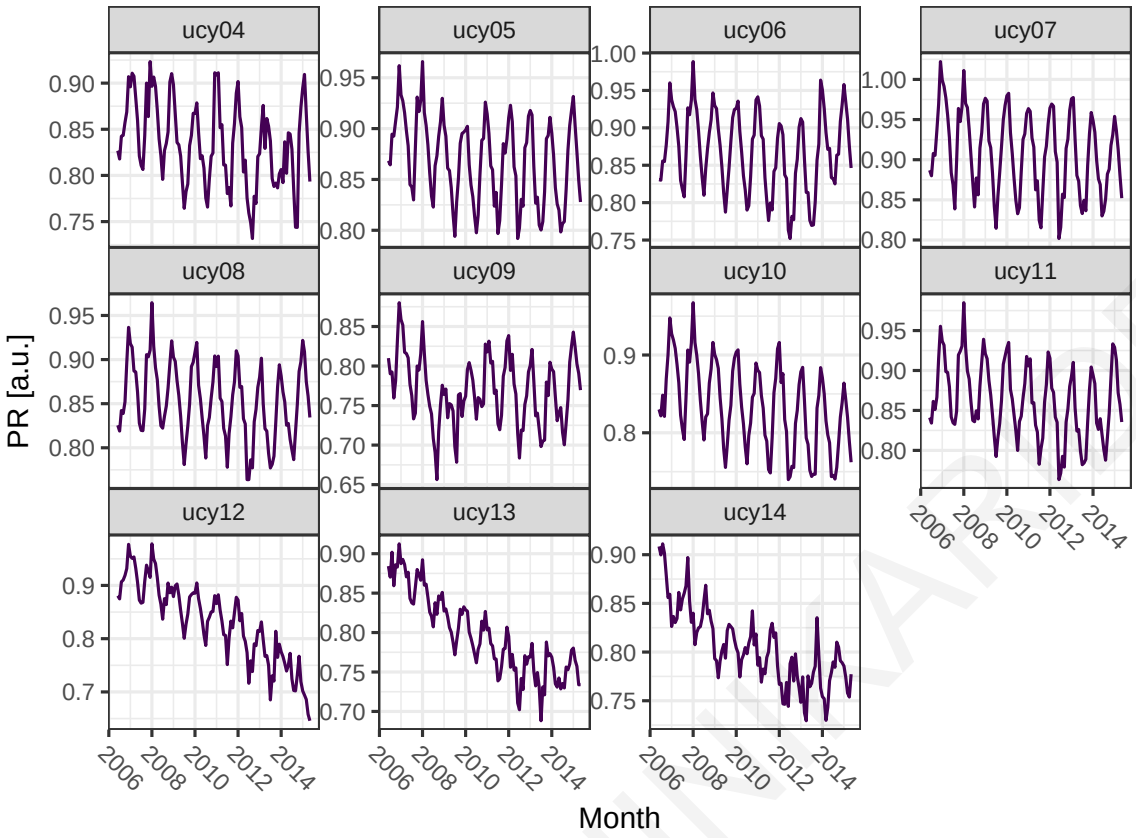


Figure 4.1: Performance Ratio (PR) of the PV systems under study at the University of Cyprus, from June 2006 to May 2015.

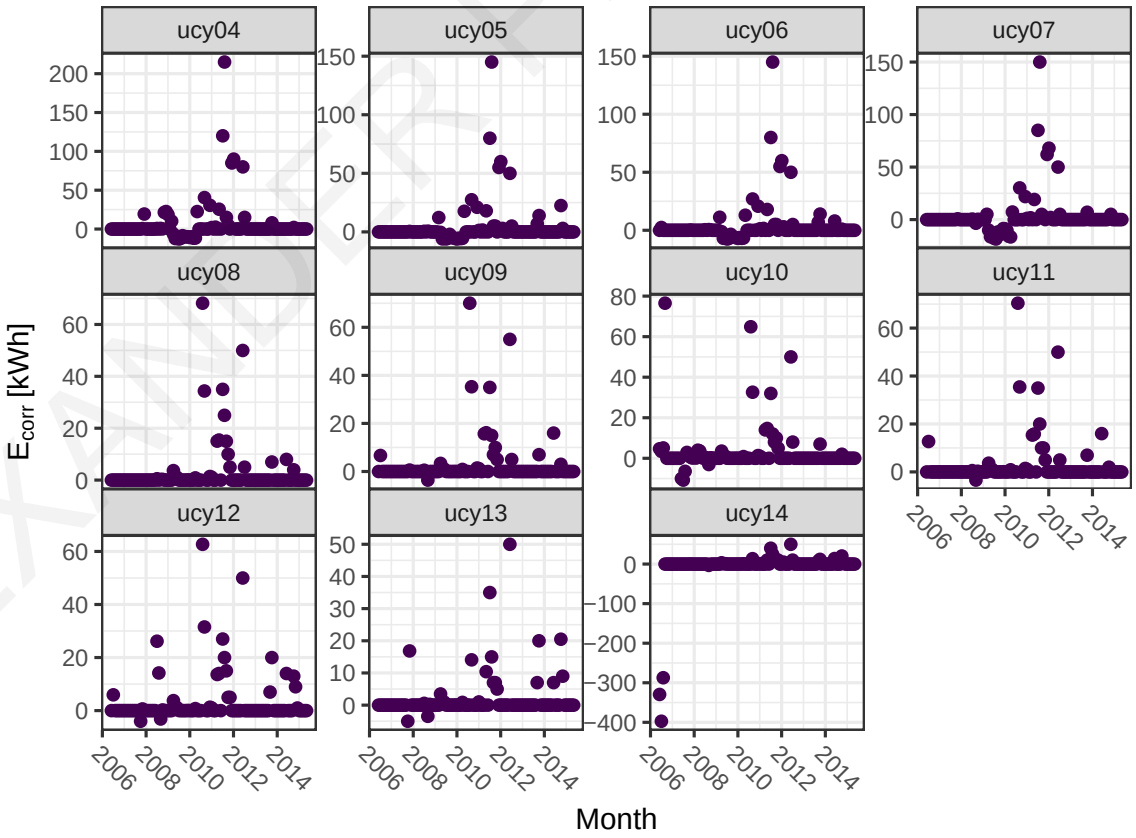
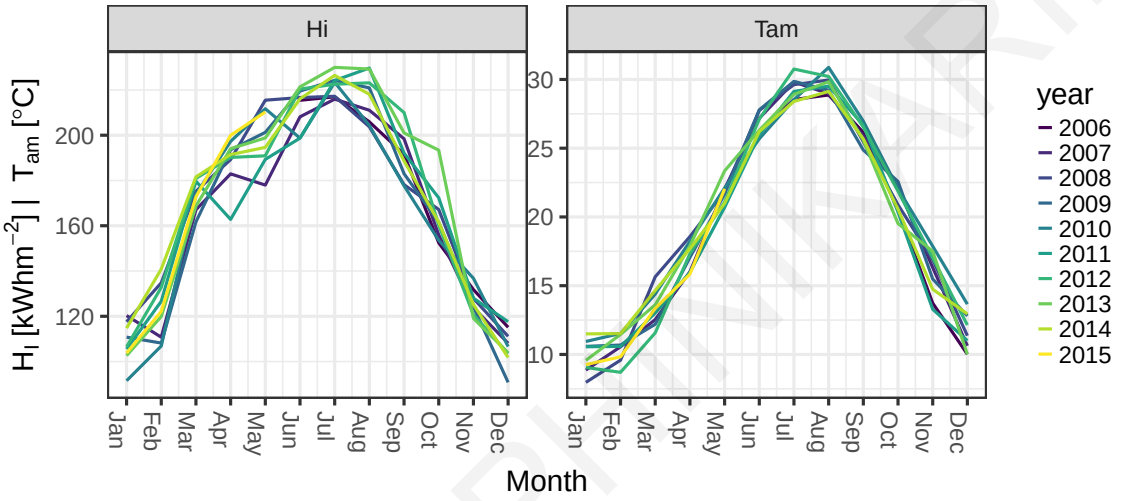


Figure 4.2: Monthly energy yield corrections, estimated empirically.

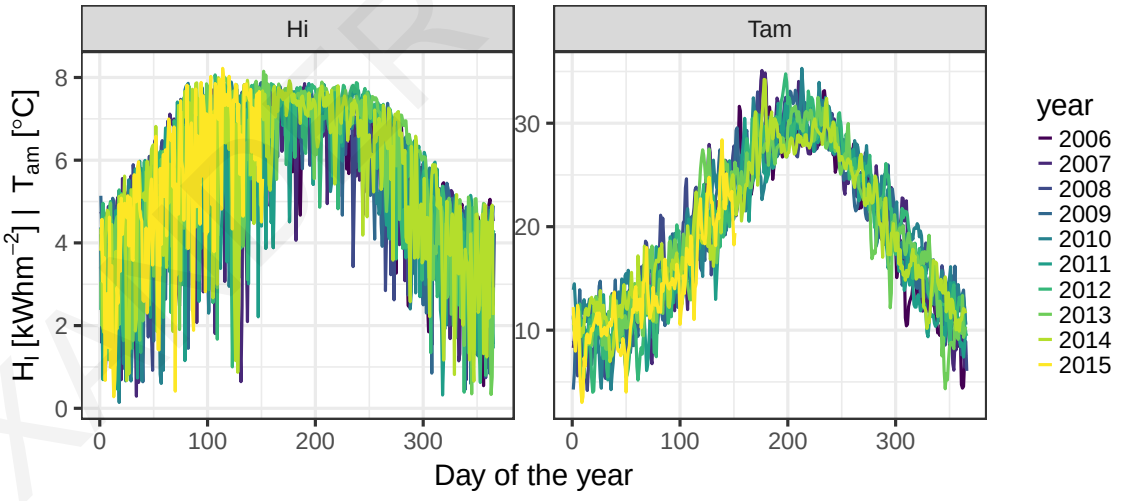
required, to be able to compute the sun's trajectory, the angle of incidence, θ_{AOI} and the sunrise and sunset times.

4.2 Prevailing meteorological conditions

The prevailing meteorological conditions at the PV Technology test site were typical for the eastern Mediterranean region, with mostly sunny days, as shown in Fig. 4.3a and high ambient temperatures during the day, as shown in Fig. 4.3b. From day 75 until 244 of each year, the irradiance at the POA reached its highest point (8 kW h/m^2.)



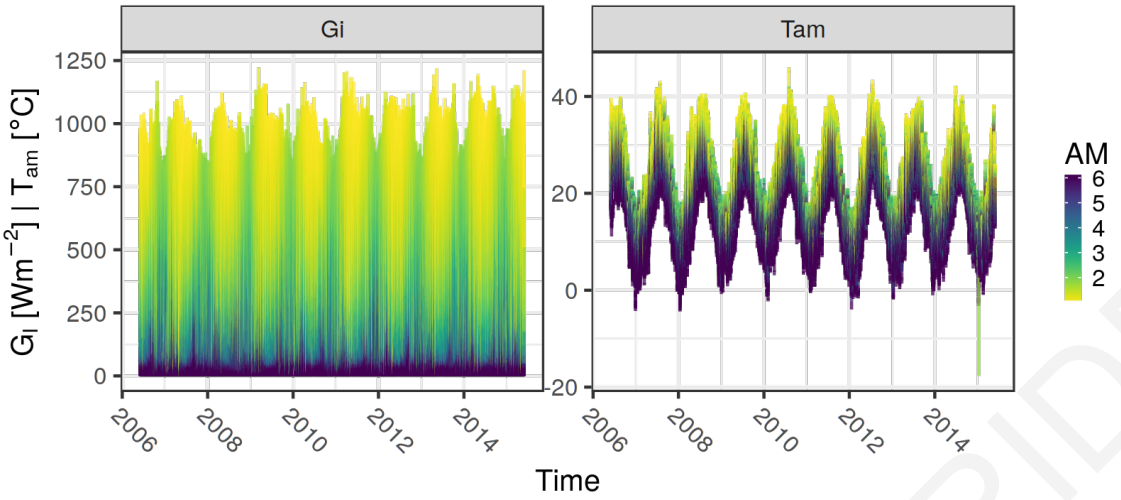
(a)



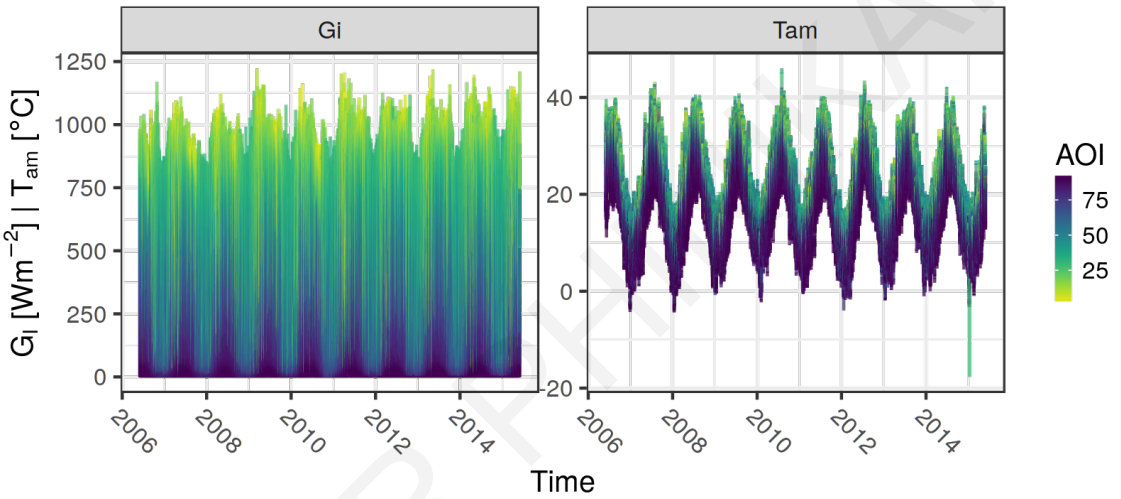
(b)

Figure 4.3: (a) Monthly and (b) daily irradiance measured with a Kipp & Zonen CM21 pyranometer on the POA and ambient temperature.

The highest measured G_I and T_{am} were observed at instances of low AM and low θ_{AOI}. This can be seen in Fig. 4.4a and Fig. 4.4b which shows plots of the fifteen-minute average G_I and T_{am} at different levels of AM and θ_{AOI}.



(a) AM Interaction



(b) AOI Interaction

Figure 4.4: Fifteen-minute average G_I measured on the POA and ambient temperature showing interaction with (a) Air Mass, and (b) Angle of Incidence.

4.3 Exploratory data analysis

4.3.1 Distribution of data

The distribution of the data can be checked graphically (e.g. histogram, quantile-quantile plot), or formally. An initial assessment of the underlying distribution, is shown in the histograms for P_A in Fig. 4.5 and for the instantaneous performance ratio, iPR (which is defined in Sec. 5.2) in Fig. 4.6. The underlying data were selected for daylight periods only, to avoid inflating the probability density function (PDF) with zeroes. From the histograms it can be seen that the P_A distribution resembled a zero-inflated variable, than a normally distributed one. The iPR distribution was also non-normal.

Formal statistical tests measure the uncertainty of the H_0 and report a p-value, which is the probability of rejecting the H_0, given that it is true. In this case, the H_0 assumes that the data is normally distributed. If the p-value is greater than α, it means that there

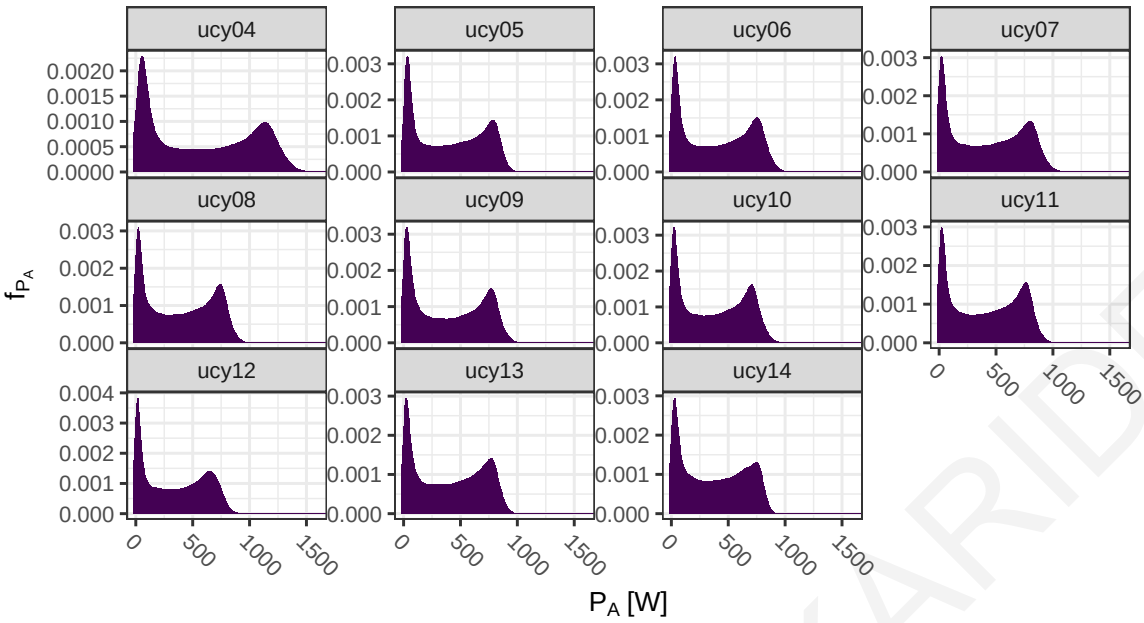


Figure 4.5: Histograms of the P_A of the PV systems under study, restricted to daylight only.

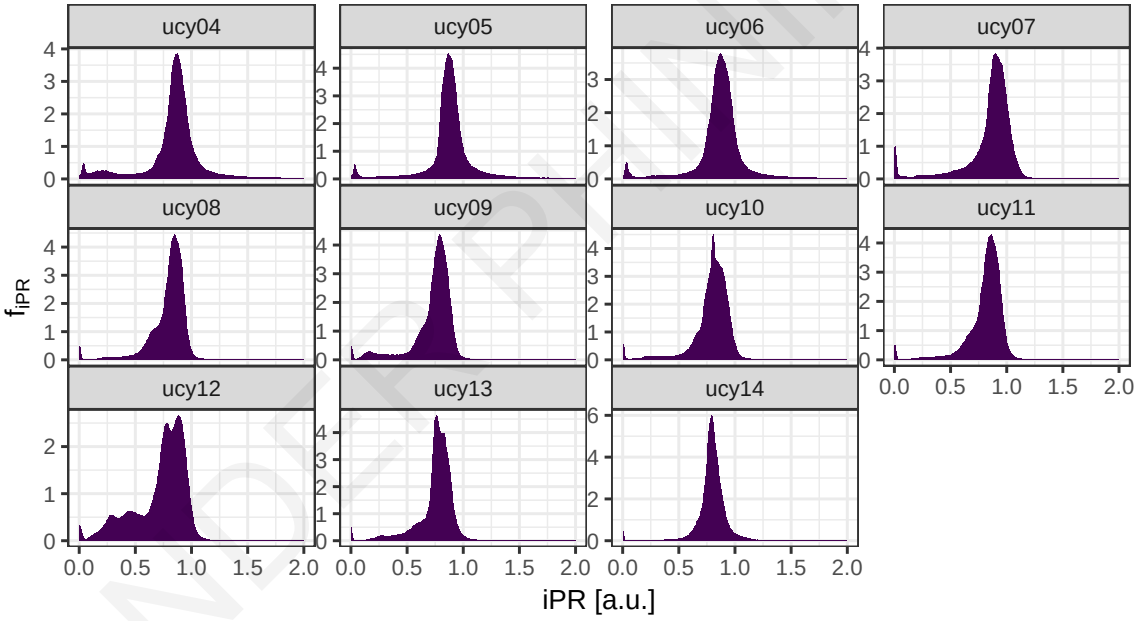


Figure 4.6: Histograms of the *iPR* of the PV systems under study, restricted to daylight only.

is no evidence to reject the H_0. Otherwise the H_0 must be rejected. Therefore, statistical significance is achieved when the p-value is less than the defined α.

To test data normality in a formal way, two well studied test statistics were used: 1) the Shapiro-Wilk test [144], and 2) the Kolmogorov-Smirnov test [145, 146]. These two non-parametric tests were used to assess the normality of the data. Non-parametric tests do not specify a distribution a-priori, therefore do not rely on the assumption of a known population PDF for their validity [147].

The Shapiro-Wilk H_0 assumes normality, therefore a low p-value of the Shapiro-Wilk test statistic rejects the H_0. The p-values are shown below for the P_A in Fig. 4.7 and the *iPR* (which is defined in Sec. 5.2) in Fig. 4.8. The y-axis was square-root transformed for

better visual inspection. It can be seen that the p-value of W for the P_A was below the

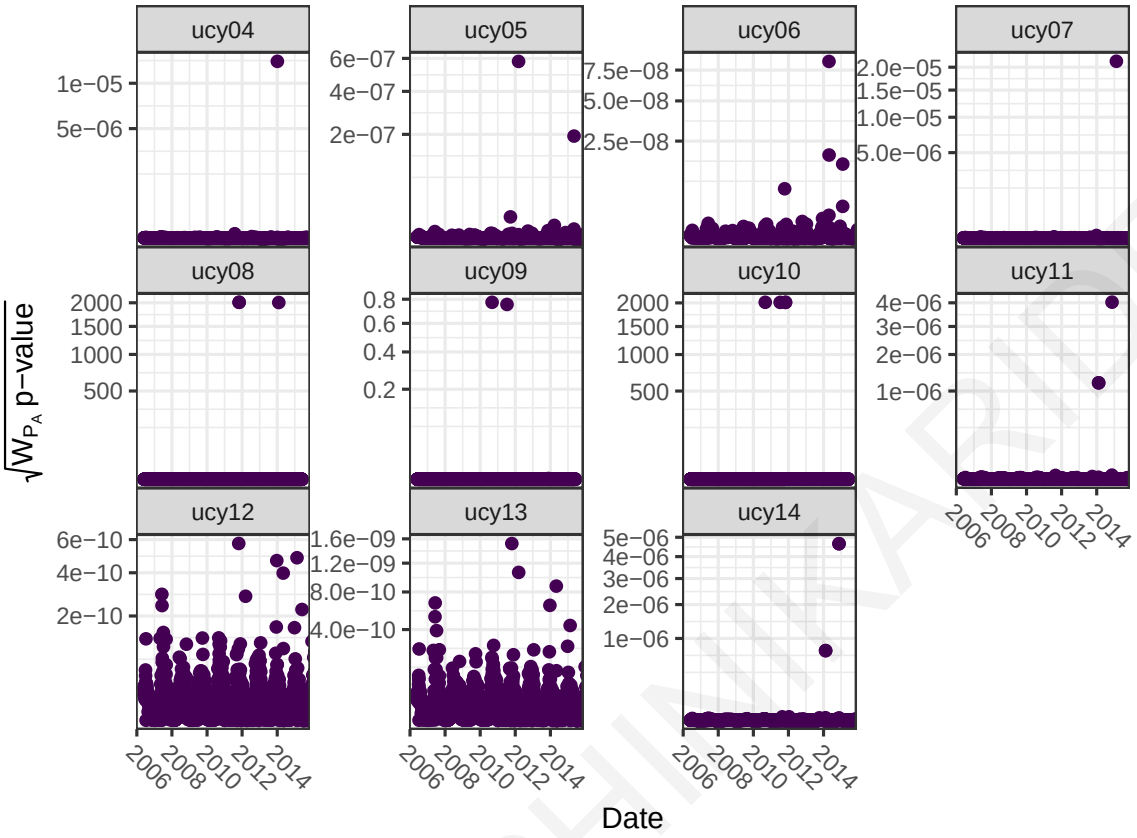


Figure 4.7: Square roots of daily p-values of the Shapiro-Wilk test statistic on P_A.

significance level $\alpha = 0.05$ across all PV systems, whereas for the *iPR*, and especially for ucy12, there were many instances where the H_0 could not be rejected.

The Kolmogorov-Smirnov test estimates the distance between the sample cumulative distribution function (CDF) and the reference CDF. The H_0 is that the sample is drawn from the reference distribution and again, a low p-value rejects the H_0. The p-values of the Kolmogorov–Smirnov test statistic are shown in Fig. 4.9 for the P_A and in Fig. 4.10 for the *iPR*. The Kolmogorov-Smirnov test produced similar results to the Shapiro-Wilk for the P_A but different for the *iPR*. Whereas the W p-value indicated some notion of normality, the D_n was stricter and its p-values rejected the notion of normality.

4.3.2 Correlation of the covariates

The Pearson correlation coefficient, $\rho_{X,Y}$, is a measure of the strength and direction of association that exists between two continuous variables, X and Y. Its value can range from −1 for a perfect negative linear relationship to +1 for a perfect positive linear relationship. A value of 0 indicates no relationship between two variables. The Pearson correlation coefficient is obtained by dividing the covariance of the two variables, $E[(X - \bar{x})(Y - \bar{y})]$, by

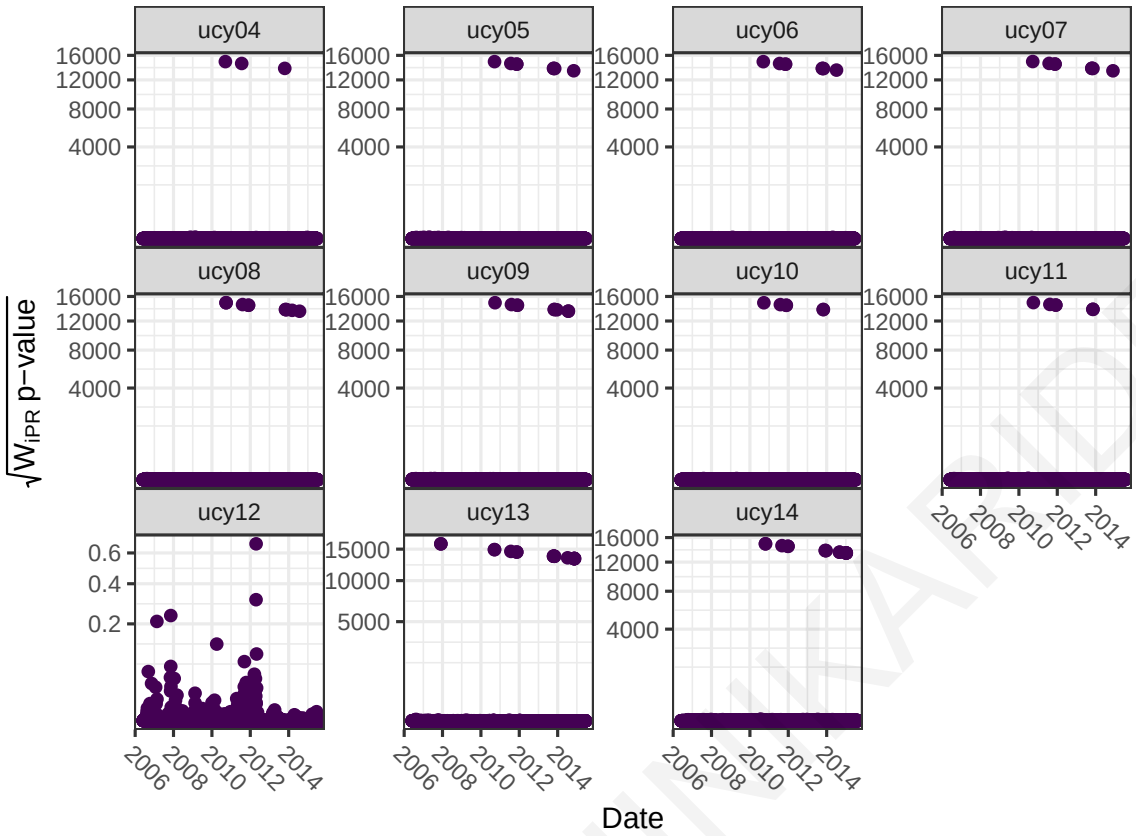


Figure 4.8: Square roots of daily p-values of the Shapiro-Wilk test statistic on *iPR*.

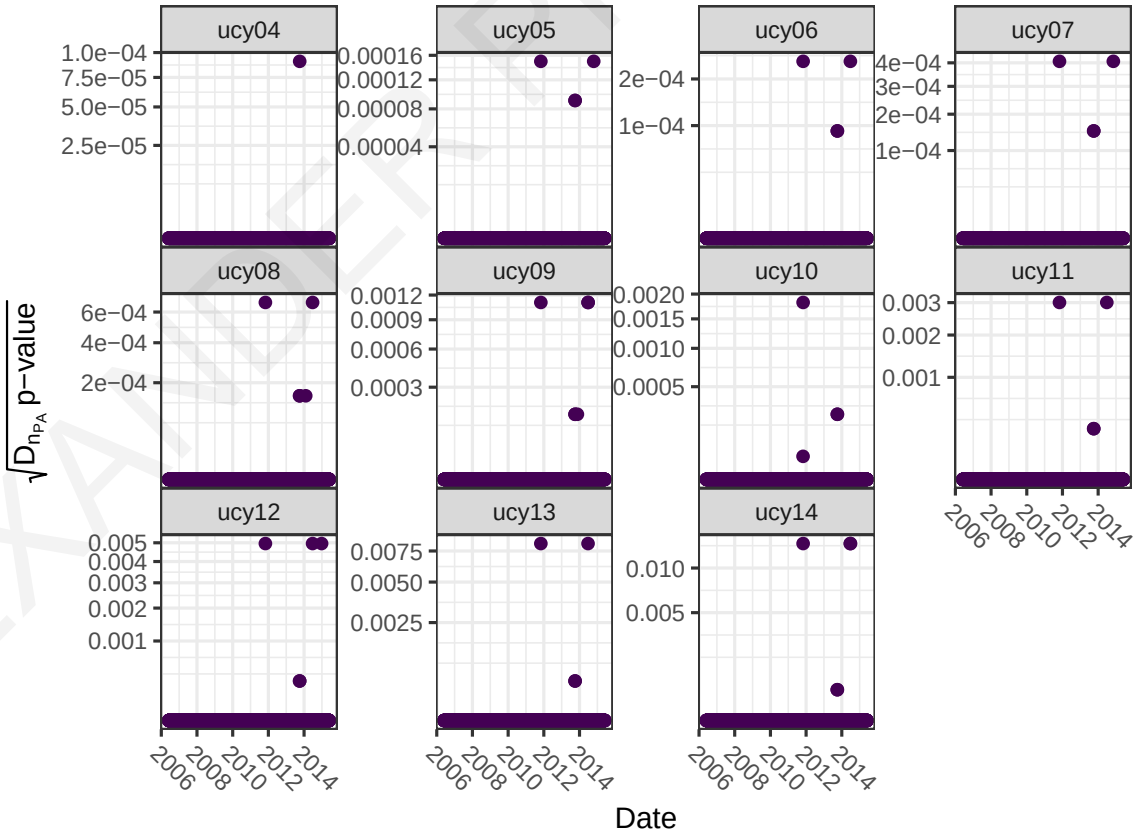


Figure 4.9: Square roots of daily p-values of the Kolmogorov–Smirnov test statistic on P_A.

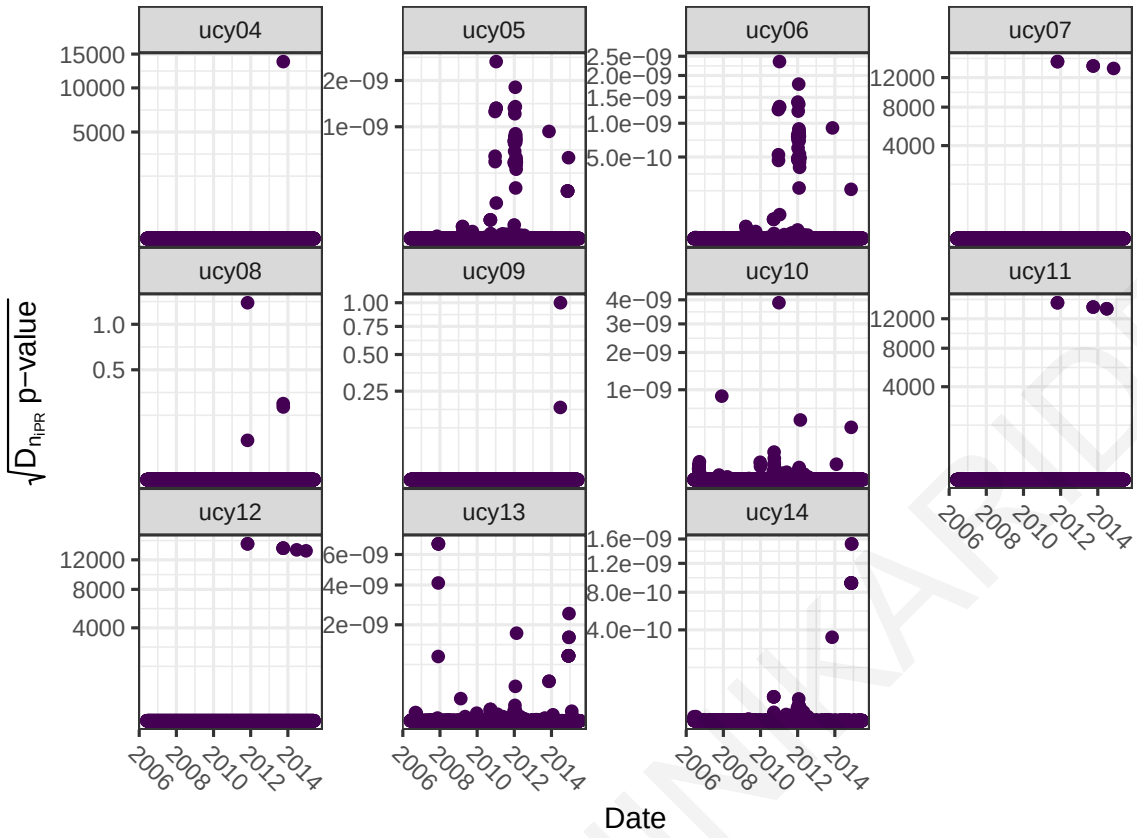


Figure 4.10: Square roots of daily p-values of the Kolmogorov–Smirnov test statistic on *iPR*.

the product of their standard deviations, as in Eq. 4.1.

$$\rho_{X,Y} = \frac{E[(X - \bar{x})(Y - \bar{y})]}{\sigma_x \sigma_y} \tag{4.1}$$

Near perfect correlation was observed between P_A and G_I, with a 0.97 correlation coefficient, for all the PV systems under test, as shown in Fig. 4.11. As expected, other variables also had strong correlation to the P_A, namely the θ_{AOI} and the T_m. These independent variables were also strongly correlated to the G_I and if they were included in the data model, confounding would occur.

The relationship between P_A as a function of G_I and T_{am} was also explored. In Fig 4.12 it can be seen that most of the variability in P_A/G_I could be found at lower T_{am}. From the figure, distinct linear relationship could be distinguished, although for ucy10 and ucy12, multiple slopes could be distinguished.

Similarly, the relationship between P_A as a function of T_m and G_I is shown in Fig. 4.13. From the figure, a strictly linear relationship was difficult to ascertain. For this reason, which is additionally validated by other studies in the literature [148, 149, 71, 150, 151], this work does not use T_m for correcting P_A measurements.

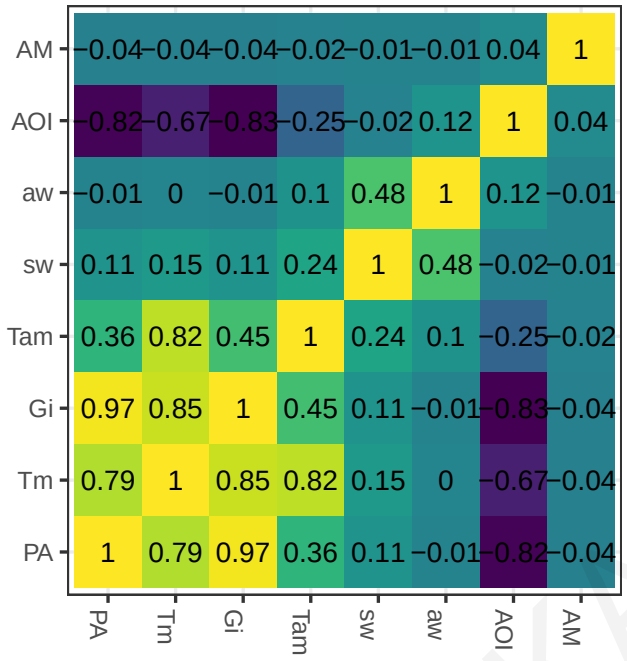


Figure 4.11: Pair-wise Pearson correlation coefficient between pairs of variables.

4.4 Data qualification

The guidelines in [36] were used to ensure that the measurement data did not contain systematic errors and that they were measured accurately, which was checked via the monthly calibration campaigns described in Sec. 3.1.3 and the periodic accredited calibration.

The data qualification flowchart is shown in Fig. 4.14. The first step in the analysis of measurement data was the removal of invalid values and the detection of missing timestamps. To assess the amount of data missing from the complete fifteen-minute data set, the data were aligned to a regular grid and the gaps were filled with `NA` values. Invalid data was defined as points with invalid values, such as `NaN`, `Inv`, `Inf` and measurement data that were outside the measuring range of the instruments. These out-of-range measurements consisted of a few instances of spikes and stuck values which recorded power generation during the night.

The above procedure was fully automated, to minimize human error, and made use of the sunrise and sunset times calculated as described in Sec. 3.2.1. Fig. 4.15 shows the amount of missing data points from the raw fifteen-minute data set, annually and monthly, after performing data qualification. Two areas of about 300 missing data points from the fifteen-minute data set were found in 2013 and a lesser amount in 2011. The events in 2013 corresponded to a database corruption issue due to a power cut and inability to recover the data points for specific days. The events in 2011 corresponded to the rolling black-outs issued by the Electricity Authority of Cyprus, following catastrophic damage to one of its power plants.

Lastly, the only manual manipulation on the data was a correction of data-logger misconfiguration during the first few months of operation of the ucy14 PV system back in June

2006, by applying an estimated calibration factor to the V_A and I_A measurements.

4.5 On-line analysis

In this approach, the evaluation is based on analysis of field measurement data and consists of a pipeline of data filtering [65, 80], outlier identification [83, 152], dimensionality reduction [91, 153], seasonal decomposition [154] and trend modelling [103, 155, 156] procedures.

Each PV system and its metadata were treated as members of list of PV systems under study. This abstracted the analysis from the actual PV system specifics and allowed the development of efficient, reproducible and unsupervised data analysis. The split-apply-combine paradigm [157] was used extensively to shape the unsupervised R_{D_E} estimation approach presented in this dissertation. Following this paradigm, functions used to manipulate data were created to be as generic as possible by accepting general data structures, performing the specific transformations and returning the same type of data structures. In *R*, this functionality is provided by the *dplyr* [158] and *purrr* [159] family of tools. In this work, each member of the list was treated as a separate entity for analysis. The generic functions were applied on members of the list using the iterative split-apply-combine paradigm. The use of the iterative approach made large parts of the analysis open to explicit parallelism which was extensively used throughout the analysis.

4.6 Challenges

As described in Sec. 4.3, PV system and meteorological measurements did not follow a normal distribution. This posed difficulties in applying statistical inference tests that assume normality.

Some possible solutions to this challenge were a) to apply a transformation to the data to make them normally distributed (e.g. Box-Cox), b) use statistics that are robust or c) use non-parametric tests. In this work, the third option was preferred, to make sure that the proposed methodology could be applied on PV systems in a different climate region, under different prevailing meteorological conditions. Non-parametric tests can be especially useful with a small sample that is skewed or a sample that contains several outliers.

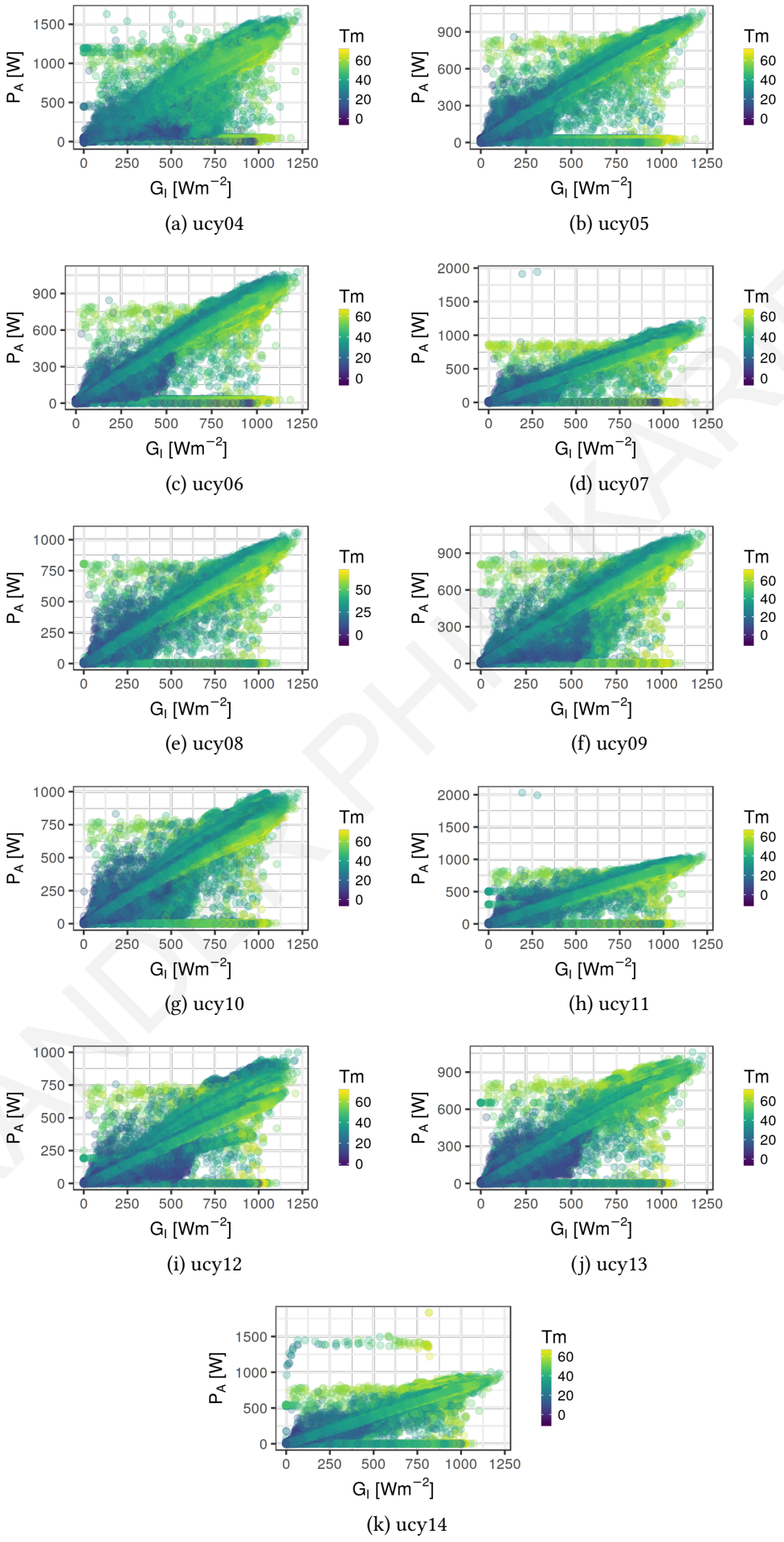
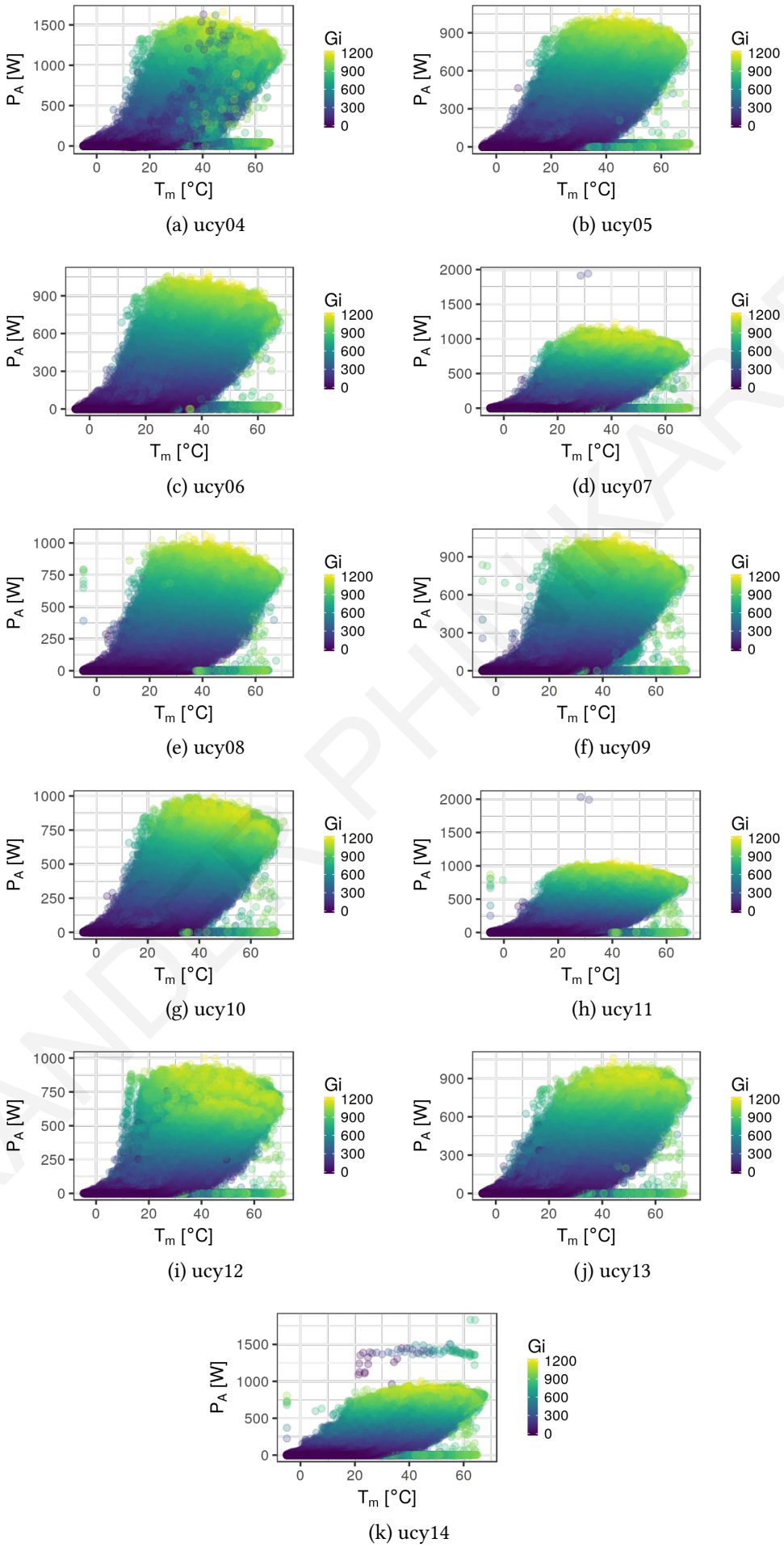


Figure 4.12: Fifteen-minute average P_A as a function of the G_I measured on the POA and interaction with T_m for the PV systems under study.



(a) ucy04 (b) ucy05

(c) ucy06 (d) ucy07

(e) ucy08 (f) ucy09

(g) ucy10 (h) ucy11

(i) ucy12 (j) ucy13

(k) ucy14

Figure 4.13: Fifteen-minute average P_A as a function of the T_m measured on the back of the modules and interaction with G_I for the PV systems under study.

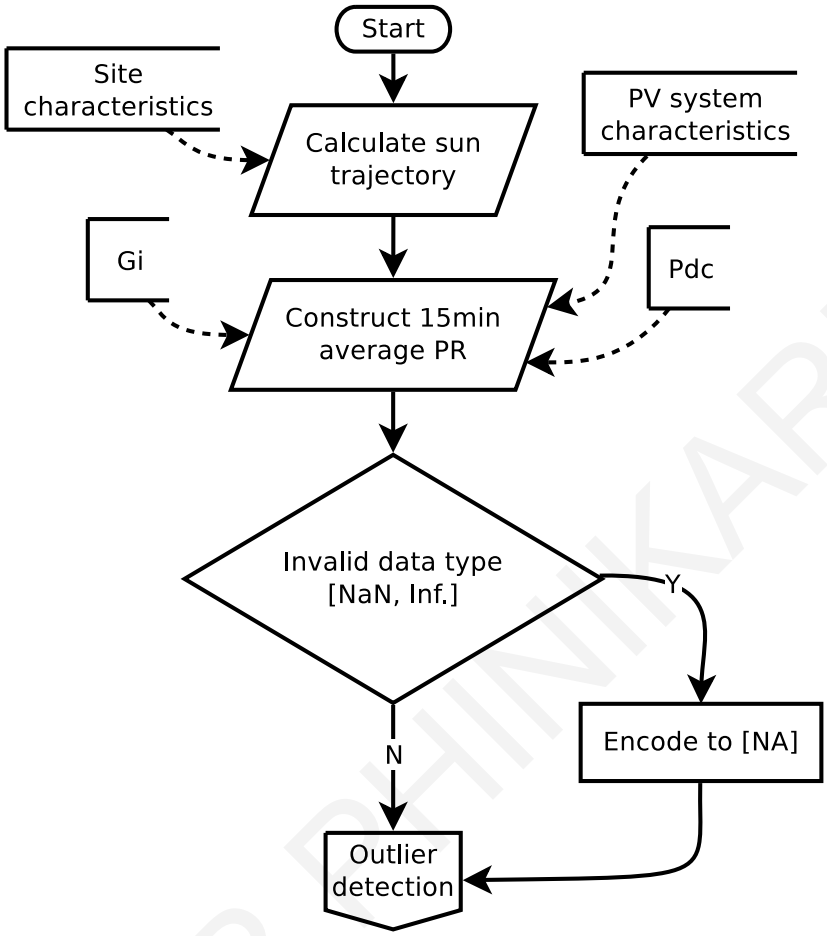


Figure 4.14: Flowchart of the data qualification procedure.

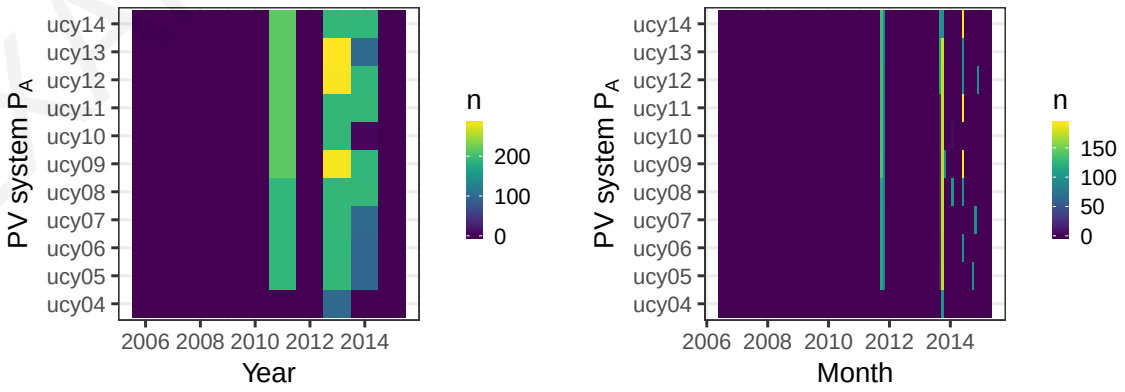


Figure 4.15: Amount of missing data points from the raw data.

Chapter 5

Detection and Treatment of Outliers

Work from this chapter has been published in [160, 161, 162]

5.1 Introduction

The subject of outlier detection in PV is very important for securing the technology and assessing power delivery in a dynamic operating environment. In the field of PV, listwise deletion, excessive filtering and discarding of un-favourable data are commonly employed in the assessment of PV performance in the field and R_{D_E} estimation studies. In this work, a rigorous statistical methodology was developed, to improve upon the existing ad hoc approaches, making use of all measurement data available for analysis.

The developed methodology treats the outlier detection task as a one-class classification problem, since the measurement data were not labelled in any way. This meant that there was no indication of catastrophic system states, only the statistics of normal operation could be discovered. Abnormal operation, during e.g. shading, cloud coverage, system and grid faults was assumed to be intermittent and had to be identified from the data. Even though instances of total system outages were logged for the PV systems under study, the assumption that there was no prior information on the state of the system was a reasonable one to make since since, in the real world, PV systems are very rarely maintained and monitored the same way as in a research-grade infrastructure.

Alternatively, in a supervised learning fashion, a theoretical model of PV performance could be constructed to provide a labelled data set. This would have the advantage of being able to compare the expected to the actual performance, assuming that all extrinsic factors could be modelled or measured (e.g. PV panel degradation, shading, clouds, soiling, precipitation, PV panel cooling from the wind, interactions with fauna and flora), which cannot be assumed to be true in every case. Typical model-based fault detection for PV performance would thus require investing in sensor hardware or specialized equipment that can perform IV characterization in situ, which could provide much more information than single point measurements.

The main requirements that defined the developed methodology were the following:

- Continuous updating of the underlying detection task as new data becomes available.
- Minimization of false positives.
- Ignore seasonal variability.
- Robust so that it works on data that is not normally distributed.
- Computationally efficient so that it converges quickly.

5.2 Data transformation

To be able to detect an outlier in time series of P_A, a relatively stable metric needed to be constructed, which was later assessed for the presence of outliers (or anomalies). One option was to transform the P_A, and construct a performance metric that aided the detection of outlier points. Since, by definition, the P_A was perfectly linearly correlated to the G_I, their ratio normalized by the PV module's P_{nom} captured most of the PV system losses. This is evident in the Pearson correlation matrix plot in Fig. 4.11, which plots the correlation between the variables in the data set and in Fig. 4.12 which proves the linear relationship.

Therefore, for the purpose of this analysis, the new normalized and scaled metric was defined in order to generalize the proposed outlier identification procedure across all PV system scales and irradiance levels. The metric was defined by P_A measurements normalized to the array P_{nom} and linearly extrapolated to $G_{STC} = 1000\,\mathrm{W/m^2}$, according to Eq. 5.1.

$$iPR = \frac{P_A}{P_{A_{nom}} \frac{G_I}{G_{STC}}} \tag{5.1}$$

The metric was defined as the iPR, and is closely related to the PR. The main difference was that the PR is formally defined for daily or monthly or annual aggregates [36], whereas the iPR is a static metric.

This means that the fraction $\frac{P_A}{G_I}$ should essentially represent a straight line, in the absence of outliers. The ratio $\frac{P_A}{G_I}$ makes sense as a data transformation from a physical point of view, as the effects of solar variations and the 11-year solar cycle [163] would be factored in the analysis.

Under ideal conditions, the iPR should follow a straight horizontal line around 1. Under real conditions, this ratio is affected by factors such as the effect of temperature on the PV panel voltage and current, reflection losses due to the angle of incidence of G_I and front PV panel surface soiling, recombination, physical defects in the PV panel, fast moving clouds and other secondary factors.

5.3 Outlier detection

Throughout the years of field exposure it was observed that, under some circumstances, some of the arrays under study were partially shaded in the early morning or in the late afternoon due to foliage or constructions nearby. Furthermore, even though all of the modules were kept clean throughout the evaluation period, even a small amount of soiling could affect the estimation of the R_{D_E}. Other secondary effects such as abnormally high or low ambient temperature/irradiance and unpredictable cloud formations contributed to the variability of the *PR*, adding additional uncertainty to the results [164]. Due to the unpredictability of the aforementioned events and the instability of the temperature coefficients [165, 166], physical or empirical PV models were not adequate in filtering them out. Non-parametric methods were therefore chosen to try and explain the variance of the metrics and mitigate some of the uncertainty.

5.3.1 Boxplot outlier rule

Introduction

One statistical method for identifying outliers is the boxplot outlier rule, where the lower quartile (Q1, 25^{th} percentile), the median, the upper quartile (Q3, 75^{th} percentile), and the interquartile range (IQR) ($IQR = Q3 - Q1$) are used to describe the variation of the data. The boxplot method ignores the mean and standard deviation, which are influenced by extreme values (outliers). The rule can be expressed as:

$$x_1 > Q3 + 1.5IQR \bigcup x_1 < Q1 - 1.5IQR \tag{5.2}$$

The boxplot outlier rule is employed to detect values outside an estimated interval. By definition, 50 % of all measurements are within $\pm 0.5IQR$ of the median, which provides a robust measure of scale.

Bootstrapping

Bootstrapping is a popular method for providing confidence intervals and predictions that are more robust to the nature of the data [167], therefore, one of the appeals of the bootstrap is its generality [168]. Any estimate can be bootstrapped, since all that is needed are an estimate and a sampling distribution. This generality allows researchers to solve otherwise intractable problems.

Since the underlying data were not normally distributed, the non-parametric bootstrap was used to empirically estimate the sampling distribution of the boxplot statistics, without making assumptions about the form of the population and without deriving the sampling distribution explicitly. The bootstrap allowed the estimation of confidence intervals on each statistic which were used to assess the uncertainty of the proposed approach.

The essential idea of the nonparametric bootstrap is as follows: a sample of size n is drawn from among the elements of the sample $\mathcal{S}$, sampling with replacement to create the bootstrap sample $\mathcal{S}_b^*$. Next, a statistic T for each of the bootstrap samples, i.e. $T_b^* = t(\mathcal{S}_b^*)$ is computed. This procedure is repeated a large number of times, R, selecting many bootstrap samples. In this work, the number of replications were set to $R = 1000$, as common practice in other fields.

From the bootstrap samples, statistics were calculated to describe the mean, the bias and the confidence interval of the process. The mean was estimated as:

$$\bar{T}^* = \hat{E}^*(T^*) = \frac{\sum_{b=1}^{R} T_b^*}{R} \tag{5.3}$$

which was then used to estimate the bias of T, i.e. $\hat{B}^* = \bar{T}^* - T$. Similarly, the estimated bootstrap variance of T^*,

$$\widehat{Var}^*(T^*) = \frac{\sum_{b=1}^{R}(T_b^* - \bar{T}^*)^2}{R-1} \tag{5.4}$$

was used for the bootstrap estimated standard error of T:

$$\widehat{SE}^*(T^*) = \sqrt{\frac{\sum_{b=1}^{R}(T_b^* - \bar{T}^*)^2}{R-1}} \tag{5.5}$$

Finally, confidence intervals were calculated with the quantile function, which is related to the CDF. The quantile function gives the value at which the probability of a random variable is less than or equal to the given probability or significance level:

$$Q(\alpha/2) \leq CI \leq Q(1-\alpha/2) \tag{5.6}$$

where α is the significance level ($\alpha = 0.05$ in this work) and $Q(x)$ the quantile function.

The performance of the bootstrap is discussed in more detail in Appendix A.2.1.

Bootstrapped statistics

More specifically, the non-parametric balanced bootstrap with stratification was used. Points with higher weights were sampled at least once, whereas points with lower weights were sampled at most once. The weights specified to stratify the analysis weighted daytime values at 1 and night time values at 0. Night time data were labelled by calculating the sun's trajectory across the sky using Michalsky's algorithm [140], as also used in Sec. 3.2.1. In this case, night time data were weighted less than daytime data, to avoid sampling from instances when the PV systems were off.

Boxplot statistics were bootstrapped across the fifteen-minute *iPR*, day-by-day and month-by-month. Random sampling was performed separately on each PV array measurement data subset (i.e. daily and monthly subsets) to get a better sense of the individual

measures of bias.

For every PV system and for every daily and monthly subset, an index was defined to describe outlier points, which were outside the interval given in Eq. 5.2. Data points outside this interval were then replaced with the bootstrapped median which served as a robust measure of the expected iPR, to create the iPR^* dataset. iPR^* was then back-transformed into P_A^* by inverting the relationship in Eq. 5.1. This is a form of winsorizing, where the effects of extreme values are reduced.

The amount of data points identified as outliers is listed in Table 5.1 and Table 5.2 respectively for the daily and monthly outlier thresholds. As expected, the finer daily thresholds resulted in less points being classified as outliers, across all systems.

Table 5.1: Percentage of data points outside the daily bootstrapped outlier thresholds.

System	$< min$ [%]	$> max$ [%]	$< Q1$ [%]	$> Q3$ [%]
ucy04	3.122	2.296	7.258	7.152
ucy05	2.273	3.110	6.597	7.419
ucy06	1.932	2.392	6.471	7.194
ucy07	2.144	0.595	7.226	6.583
ucy08	1.807	0.422	7.439	6.521
ucy09	2.584	0.434	7.684	6.468
ucy10	2.496	0.688	7.545	6.446
ucy11	2.279	0.588	7.904	6.460
ucy12	0.786	0.270	6.446	6.213
ucy13	2.612	0.535	7.771	6.709
ucy14	1.910	2.283	7.682	7.144

Table 5.2: Percentage of data points outside the monthly bootstrapped outlier thresholds.

System	$< min$ [%]	$> max$ [%]	$< Q1$ [%]	$> Q3$ [%]
ucy04	4.148	2.699	11.368	11.380
ucy05	3.007	3.659	11.384	11.355
ucy06	2.596	2.710	11.331	11.383
ucy07	3.693	0.624	11.377	11.406
ucy08	3.293	0.467	11.389	11.361
ucy09	4.383	0.515	11.388	11.350
ucy10	3.800	0.893	11.430	11.382
ucy11	4.224	0.707	11.460	11.346
ucy12	1.830	0.233	11.320	11.351
ucy13	4.441	0.830	11.409	11.384
ucy14	3.038	2.978	11.404	11.382

Uncertainty

The uncertainty of the outlier treatment procedure proposed in this chapter was assessed by constructing a confidence interval on bootstrapped estimates. Bootstrapping was necessary in the case of PV system measurements and their transformation since the non-normality of the underlying distribution introduced bias to the confidence intervals.

The constructed 95 % confidence interval can be seen in Fig. 5.1 and Fig. 5.2 for the daily and monthly subsets of $iPR*$ respectively. The violet points represent the original data, the yellow range represents the CI of the median, the blue range represents the CI of Q1 and Q3 and the green range represents the CI of the upper and lower limits. The confidence interval in iPR was then back-transformed into u_{P_A}.

Effectiveness of the boxplot method

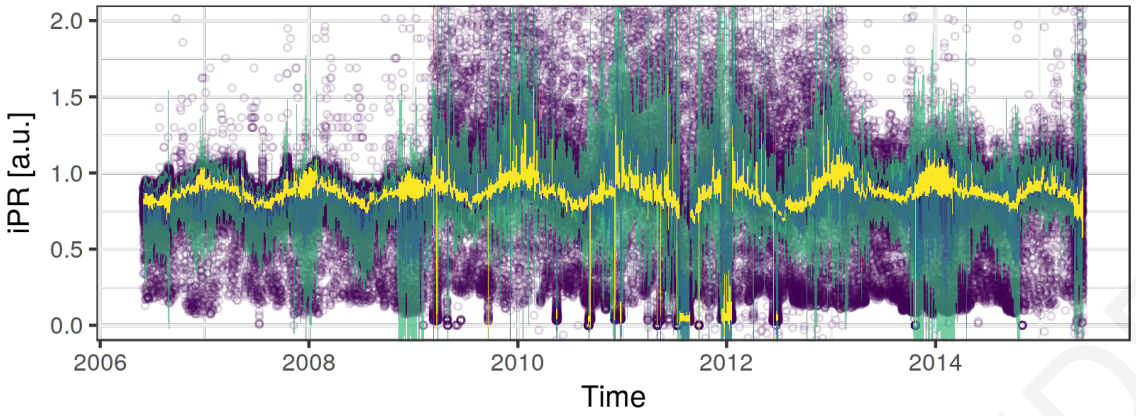
To assess the effectiveness of the proposed method, a graphical exploration of the iPR^* and the P_A^* was performed. First of all, systems that suffered from partial shading were identified and typical shading periods were plotted. Secondly, the effect of moving clouds was compared before and after applying the outlier detection method. Thirdly, the effectiveness of the daily thresholds was compared to the coarser monthly thresholds. On the one hand, the monthly thresholds were more robust to long periods of total system downtimes but provided less fidelity than the daily thresholds. On the other, the daily thresholds provided stricter confidence intervals which were not skewed by unpredictable weather conditions.

Fig. 5.3 shows the iPR^* and the corresponding P_A of the ucy07 system in spring and winter with the daily outlier thresholds, as a typical example. The graphs show the original data as a solid line, the daily confidence interval of the outlier detection method as a shaded background and the detected outliers as individual points. From these graphs, it can be seen that outlier points were successfully detected. The outliers manifested in the early morning and late afternoon be correlated to partial shading and high θ_{AOI} reflection losses. In addition, there were two outlier points detected around noon on Jan.19 which corresponded to short-term PV system outages.

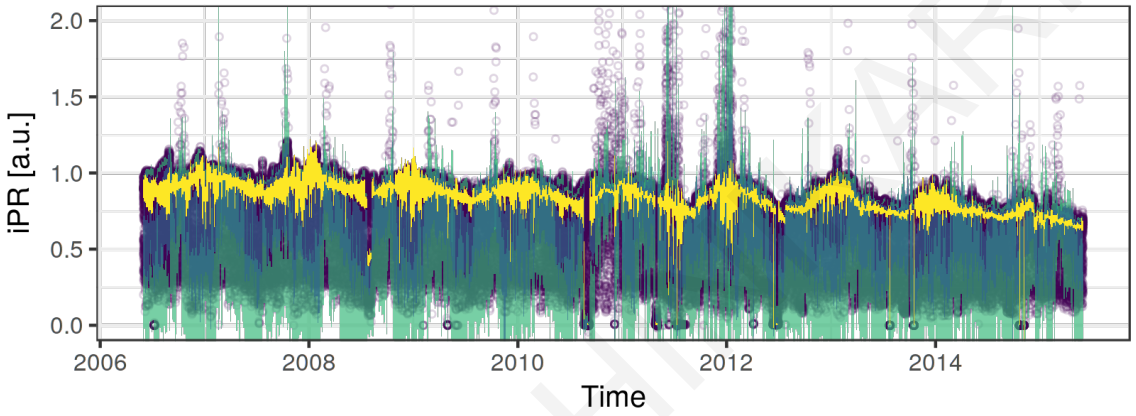
Similarly, plots of the data along with the monthly outlier thresholds can be seen in Fig. 5.4. It can be observed that the coarser monthly thresholds provided less fidelity than the daily outlier thresholds, since several instances of early morning or late afternoon shading were not flagged.

Finally, the daily and monthly methods were compared during periods of bad weather conditions. This is seen in Fig. 5.5, which shows data from the ucy14 system in winter 2015.

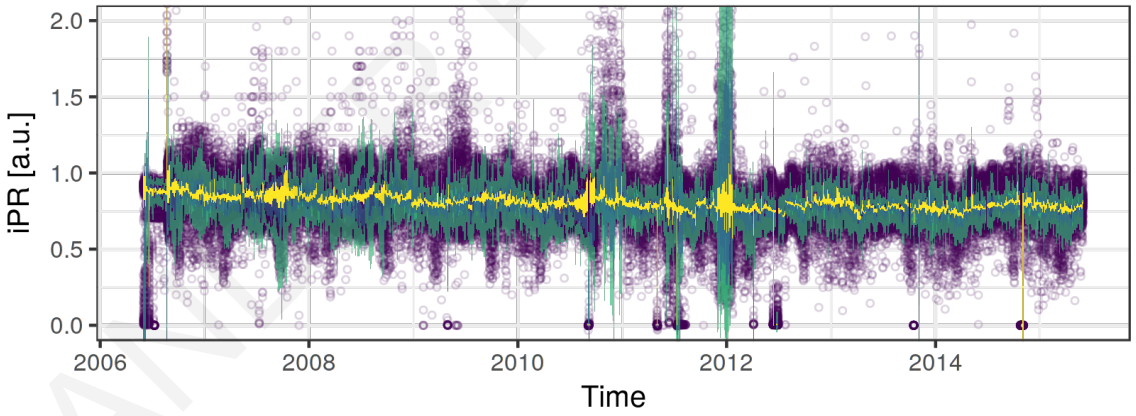
It can be concluded that although the daily bootstrapped thresholds were more sensitive to long periods of total system outages for some of the systems under study, the increased intra-daily granularity in comparison to the monthly thresholds enabled more accurate detection of abnormal losses. In an online application, the static daily / monthly thresholds could be converted to rolling daily / monthly intervals so that when new data arrive a new



(a) ucy04



(b) ucy12



(c) ucy14

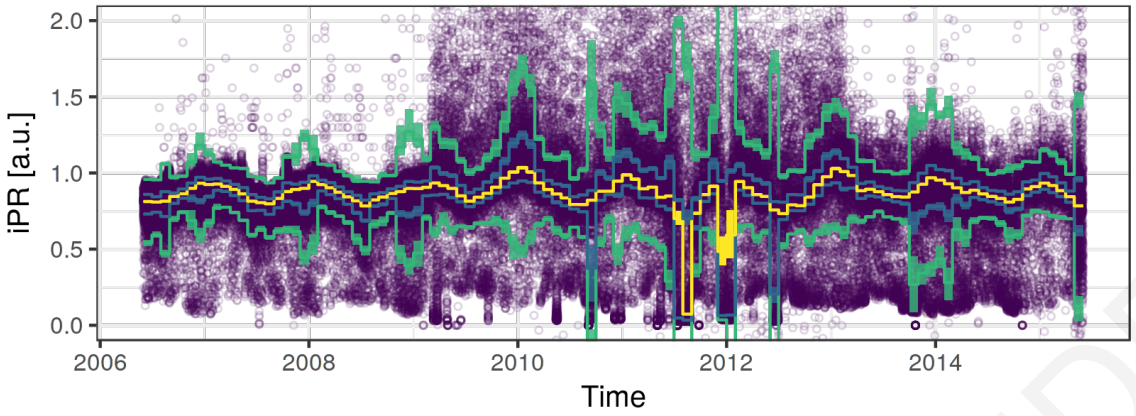
Figure 5.1: Full set of *iPR* data and the estimated daily boxplot confidence intervals.

set of thresholds would be computed on the fly.

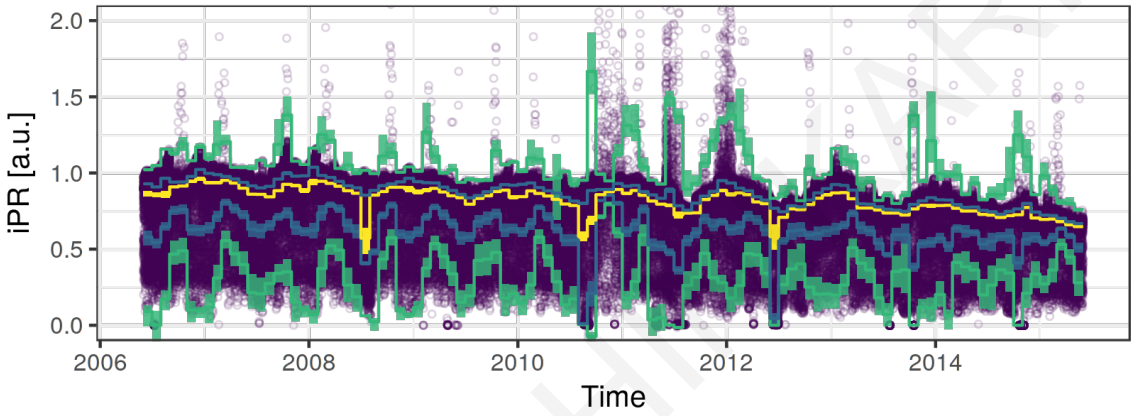
5.3.2 Principal component analysis

Introduction

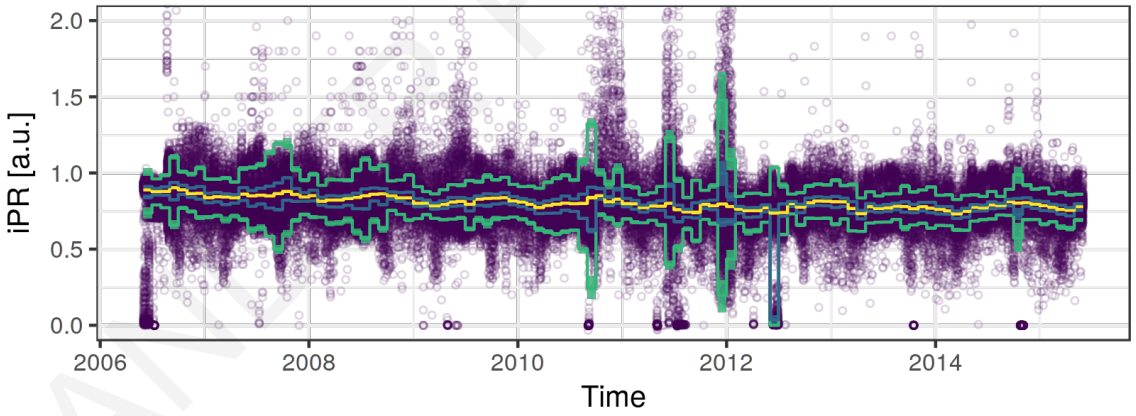
PCA is a coordinate transformation method that maps a set of data points onto new axes called the principal components. Each principal component points in the direction of maximum variance remaining in the data, given the variance already accounted for in the pre-



(a) ucy04



(b) ucy12

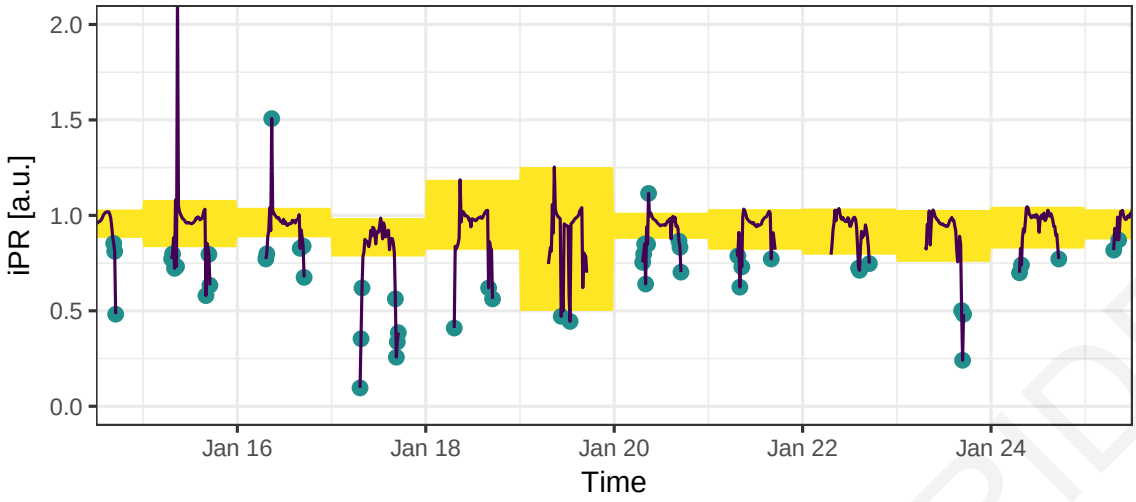


(c) ucy14

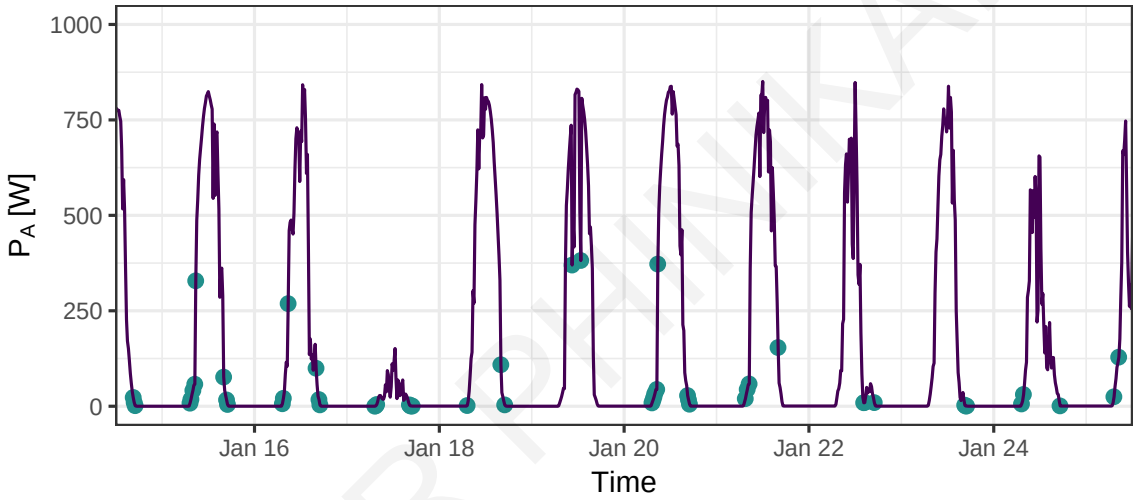
Figure 5.2: Full set of *iPR* data and the estimated monthly boxplot confidence intervals.

ceding components. As such, the first principal component is the vector that points in the direction of maximum variance. The next principal components then each capture the maximum variance among the remaining orthogonal directions. Thus, the principal axes are ordered by the amount of data variance that they capture. By examining the amount of variance captured by each principal component, it can be concluded whether the variability in the data could be captured in a space of lower dimension.

PCA, as a data transformation, dimensionality reduction, exploration, and visualization tool, does not make any assumptions. Previous studies have shown that PCA-based meth-



(a) Daily boxplot confidence intervals in iPR^* for a winter period in 2011.



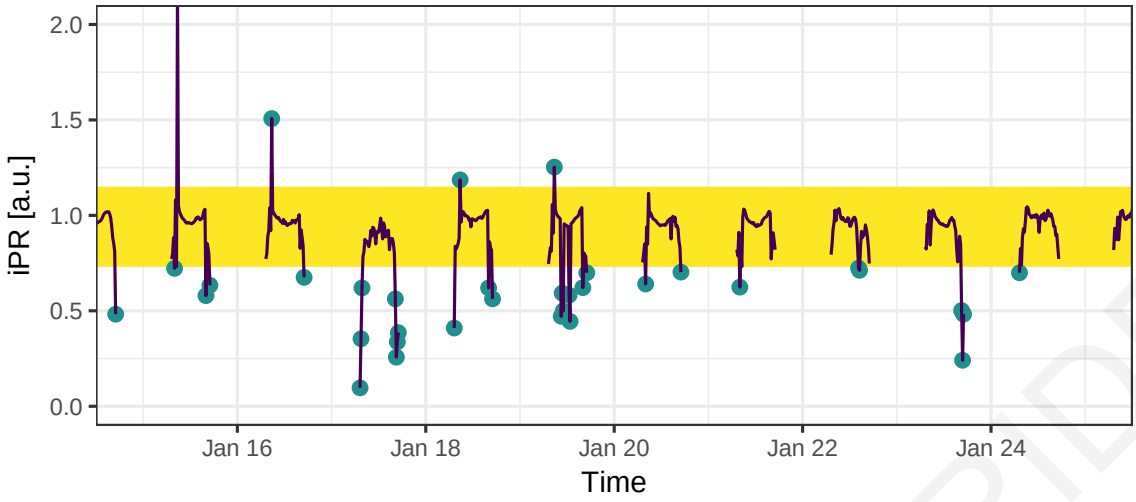
(b) P_A and detected outlier points, back-transformed from iPR^*.

Figure 5.3: Effectiveness of the boxplot outlier detection method on daily blocks of data.

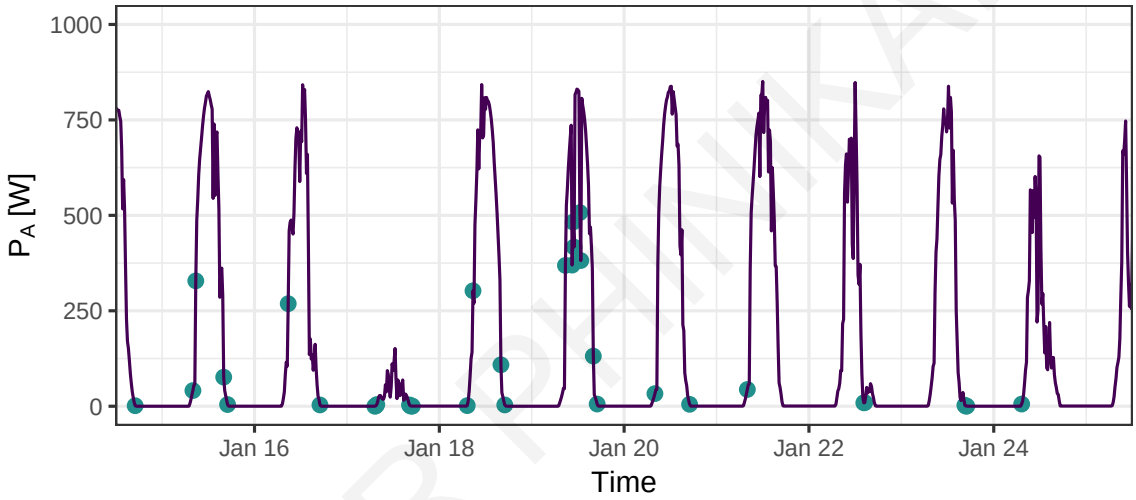
ods were successful in calculating the projection of the original variables to the principal component space, followed by the inverse projection back to the original variables [169]. When only the first principal components (i.e. the components that explain most of the variance in the data) were used for reconstruction, it was proven that the reconstruction error was low and that the variability contributed by the aforementioned outliers was removed [82]. This was due to the fact that the first principal components could explain the majority of the variance of typical values, while the rest of the principal components were associated with outlier variance and were rejected.

Application on PV measurement data

In the case of PV measurement data, it was expected that the intrinsic dimensionality would be low because of the well defined seasonal component. Higher dimensional components were expected to capture abnormal behaviour and system losses, as well as uncertainty of measurement. The PCA was applied on the iPR and iPR^* metrics, after normalizing each



(a) Monthly boxplot confidence intervals in iPR^* for a winter period in 2011.



(b) P_A and detected outlier points, back-transformed from iPR^*.

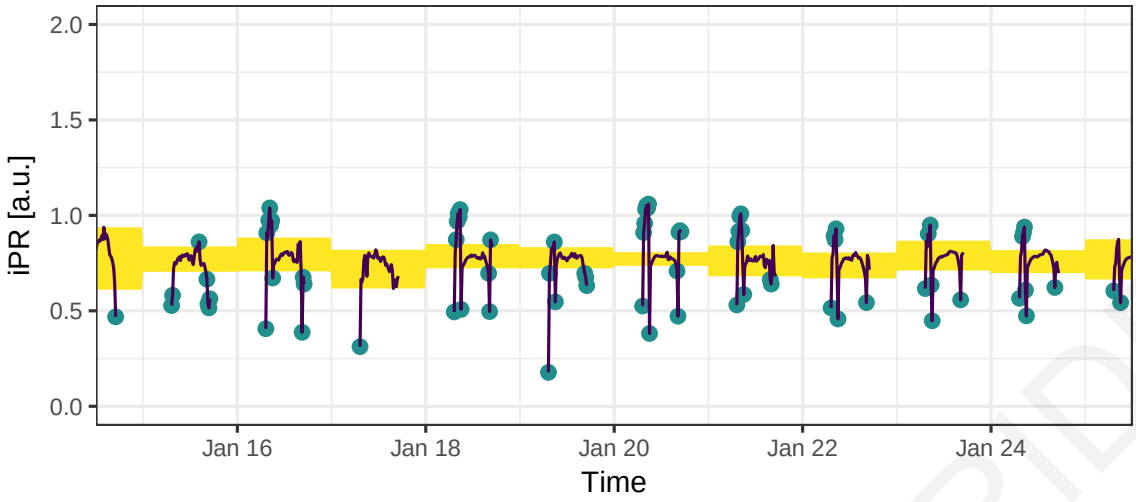
Figure 5.4: Effectiveness of the boxplot outlier detection method on monthly blocks of data.

daily vector around zero. This ensured that PCA dimensions would capture true variance, and thus avoided skewing of the results due to differences in daily means.

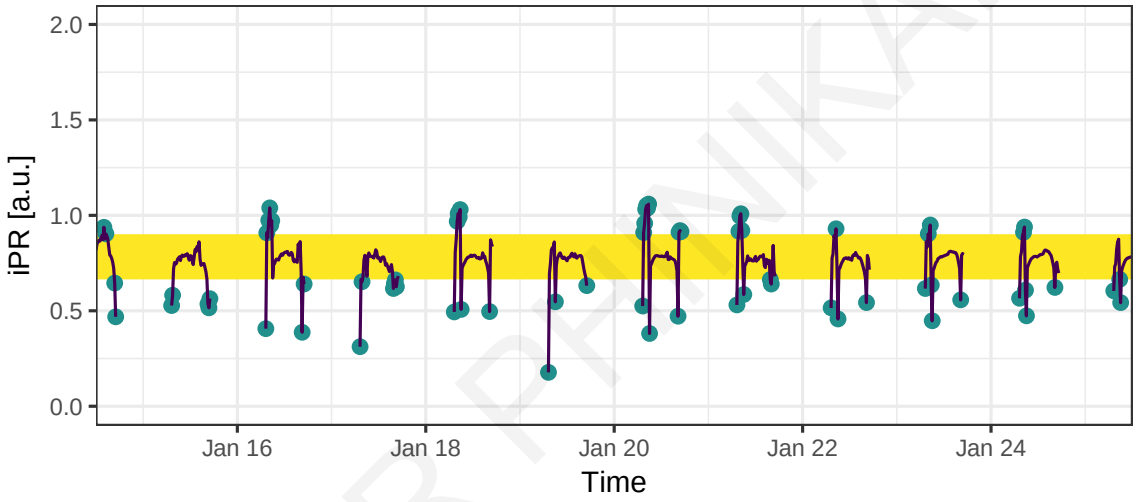
To convert the time series into the matrix **M**, each univariate series was split into vectors at the fundamental frequency (daily), i.e. for the daily seasonality, each row was a 15 min measurement point and each column was one full day of measurements. This ensured that intra-day variance was preserved and well represented.

Instances of missing data which were encoded as `NA`, as described in Sec. 4.4, were set to a sentinel value of 0, to cluster them with the rest of the anomalies, since they were undefined. In this way, a rectangular matrix was constructed. In this proposed way, the methodology became insensitive to missing data points and sensor failures since, otherwise, the PCA algorithm would fail. This made the methodology applicable in an online fashion.

The matrix, **M**, was constructed for each metric to be decomposed through PCA, consisting of 3287 vectors for the 3287 days or 9 y of the investigation period. Each vector



(a) Daily boxplot confidence intervals in iPR^* for a winter period in 2015.



(b) Monthly boxplot confidence intervals in iPR^* for a winter period in 2015.

Figure 5.5: Effectiveness of the daily and monthly boxplot outlier detection methods during instances of bad weather.

contained $\frac{60\,\text{min/h}}{15\,\text{min}} * 24\,\text{h/d} = 96/\text{d}$ elements. Each data matrix $\mathbf{M}_{i\times j}$ was defined as follows:

$$\mathbf{M}_{i,j} = \begin{pmatrix} m_{1,1} & m_{1,2} & \cdots & m_{1,j} \\ m_{2,1} & m_{2,2} & \cdots & m_{2,j} \\ \vdots & \vdots & \ddots & \vdots \\ m_{i,1} & m_{i,2} & \cdots & m_{i,j} \end{pmatrix}$$

where $i \in [1, 96]$ and $j \in [1, 3287]$ for each data set of iPR and iPR^* constructed in this work. In total, each $\mathbf{M}$ contained 3287 d × 96/d = 315552 elements and was originally rank 57.

Each vector of the matrix was then centred around zero, before PCA. Since missing data was prior-encoded to 0, the matrix was rectangular. PCA was then applied on the wide matrix of daily vectors and the number of selected principal components was chosen to explain at least 95 % of the intra-daily variance in the data. The limit is dynamically set at the point where the CDF of the variance proportions, reaches 0.95 and then rounding up.

The PCA components were obtained from the singular value decomposition (SVD) of **M**:

$$\mathbf{M} = \mathbf{U}\mathbf{D}\mathbf{V}^{\mathsf{T}} \tag{5.7}$$

where **U** is a matrix containing the left-singular vectors of **M** and **V** is a matrix containing the right-singular vectors of **M**. Lastly, **D** is a diagonal matrix of singular values calculated from the square roots of the non-zero eigenvalues of **U** and **V**.

Reconstruction

Following decomposition into their principal components, the variance explained by each component was investigated. The proportion of variance for the principal components through the transformation of *iPR* to the new axes is shown in Fig. 5.6. It can be seen that the variance was not well separated.

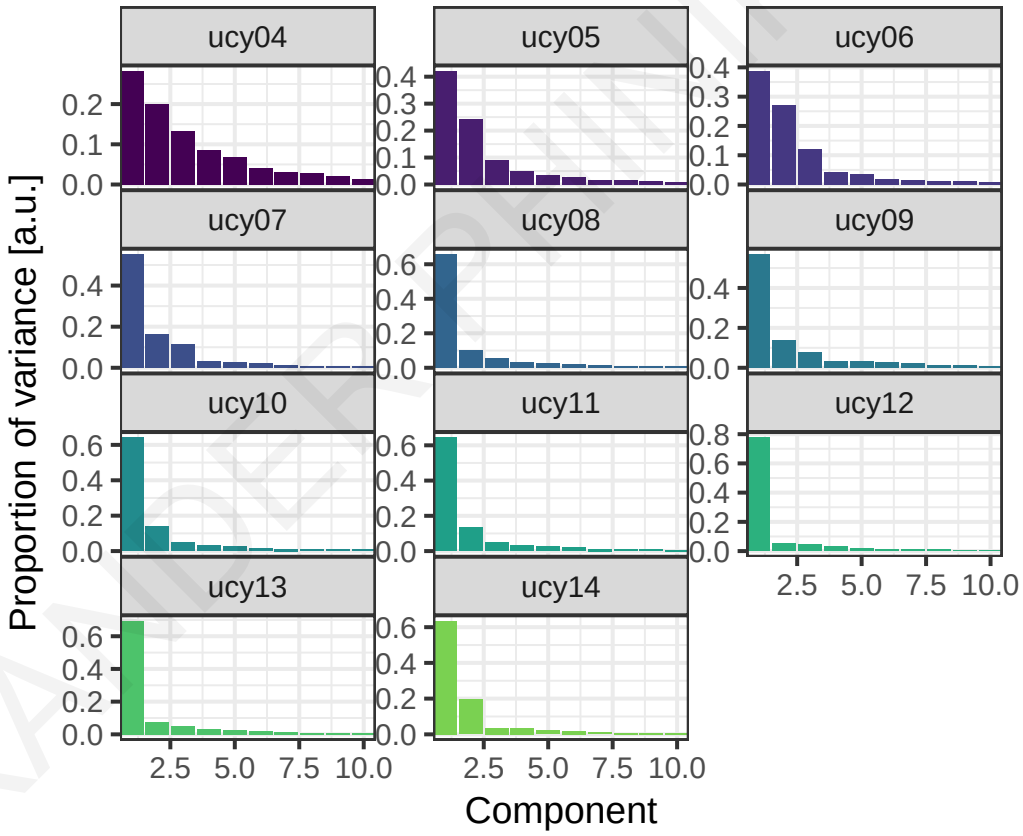


Figure 5.6: Proportion of variance of the PCA decomposed *iPR* metric.

In Fig. 5.7 the proportion of variance for the iPR^* transformed metric is shown. Using the iPR^* a much better separation of the variance into individual components was achieved. It was expected though, since the effect of outliers on the iPR^* was mitigated through the boxplot rule.

The proportion of variance explained by the first principal component is listed in Table 5.3. Very low amount of variance can be explained when using the *iPR*. A higher amount

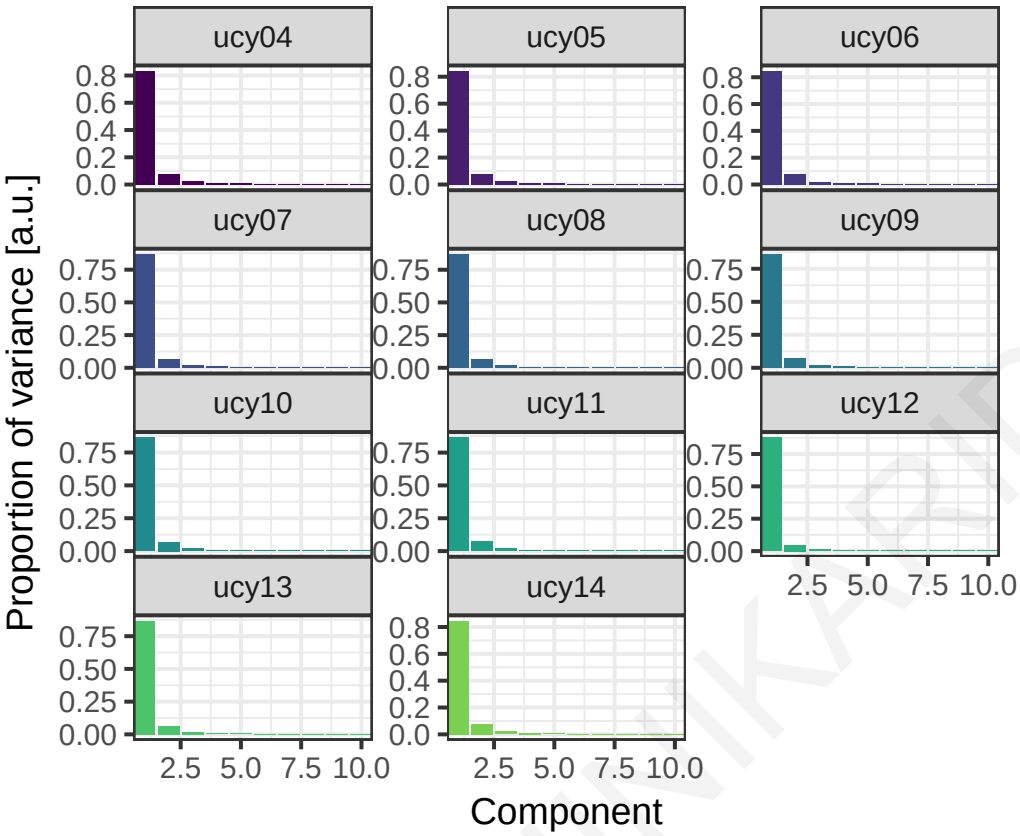


Figure 5.7: Proportion of variance of the PCA decomposed iPR^* metric.

Table 5.3: Percentage of variance explained by the first PCA component.

System	$\sigma^2_{iPRPC1}/\sigma^2_{iPR}$ [%]	$\sigma^2_{iPR^*PC1}/\sigma^2_{iPR^*}$ [%]
ucy04	28.196	83.301
ucy05	42.054	84.096
ucy06	38.716	84.289
ucy07	55.150	87.203
ucy08	65.709	87.420
ucy09	56.816	86.924
ucy10	64.067	87.076
ucy11	64.341	87.040
ucy12	77.585	88.445
ucy13	69.337	87.245
ucy14	63.481	85.395

of variance could be explained when using the iPR^*. Moreover, the amount of explained variance increased to 96 % when taking into account the first four principal components, as listed in Table 5.4.

Regarding the iPR^*, the number of selected principal components that explained at least 95 % of the intra-daily variance was 4. To assess the intra-daily variance captured by these components, the individual components were plotted on a 15 min grid, as shown in Fig. 5.8. The first component which explained most of the variance, represents what can be expected

Table 5.4: Percentage of variance explained by the first four PCA components.

System	$\sigma^2_{iPRPC1-4}/\sigma^2_{iPR}$ [%]	$\sigma^2_{iPR^*PC1-4}/\sigma^2_{iPR^*}$ [%]
ucy04	70.119	95.913
ucy05	80.339	96.325
ucy06	82.310	96.250
ucy07	86.190	96.094
ucy08	84.700	96.045
ucy09	81.596	96.067
ucy10	86.181	96.089
ucy11	85.485	96.099
ucy12	89.730	96.052
ucy13	85.345	96.393
ucy14	90.125	96.485

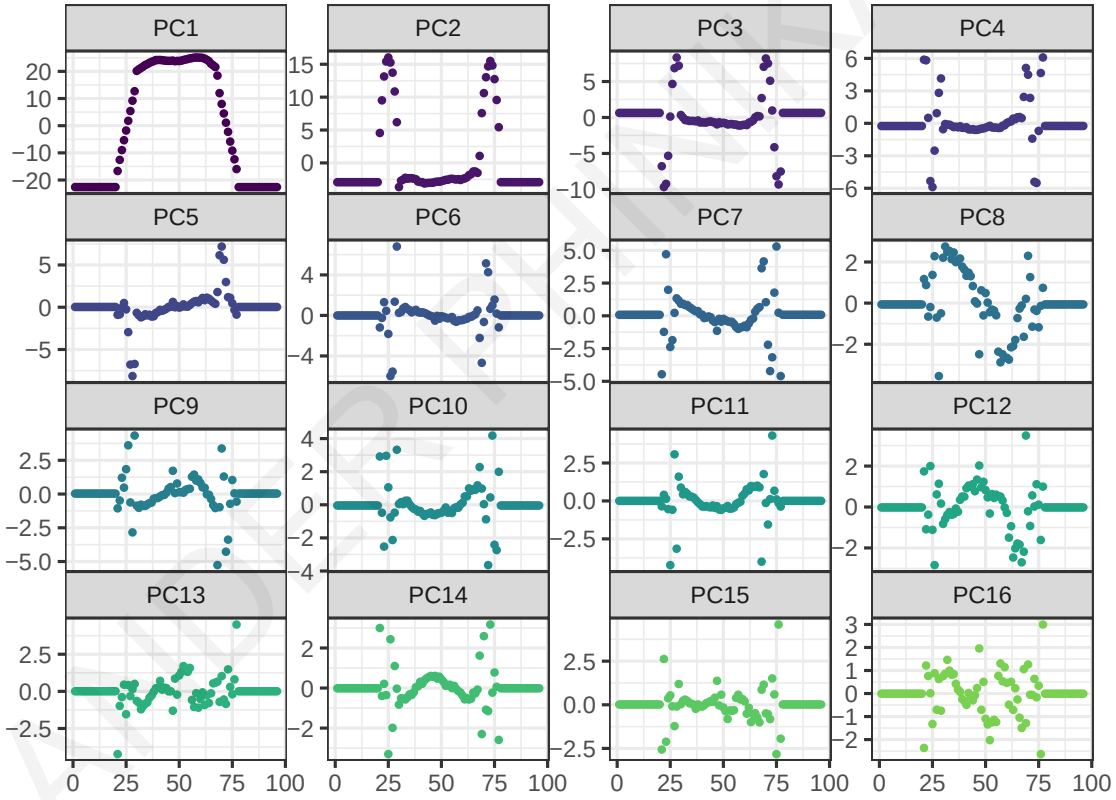


Figure 5.8: Reconstruction of the first 16 principal components of the ucy13 iPR^*.

from a PV system under near optimal operating conditions. The second principal component can be directly correlated to different sunrise and sunset times throughout a calendar year, since it's amplitude varies positively at the two extremes of the day. The variance explained by the rest of the components can be attributed to a bundled effect of partial shading and losses due to high θ_{AOI}, recombination and elevated T_{am}. After the 4^{th} component, the variance was treated as uncertainty and was discarded as it represented only 3 % to 4 % of the original amount of variance.

To assess the actual performance of the PCA, a reconstruction of the four first principal components was performed. The inverse projection of the first 4 principal components of iPR^* was used to create the iPR^*_{PCA} and the back-transformation to $P^*_{A_{PCA}}$ is shown in Fig. 5.9 for the ucy13 system as a typical example.

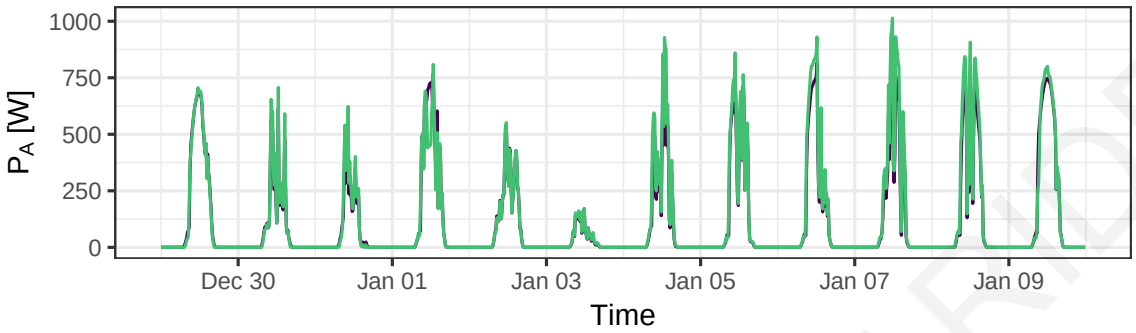


Figure 5.9: Back-transformed $P^*_{A_{PCA}}$ from the principal components of iPR^*.

It can be concluded that by discarding the components whose variance could be attributed to intermittent behaviour and then reconstructing a reduced version of the original variable, the iPR^*_{PCA} was successfully mapped to a lower dimension.

Suitability for degradation studies

To assess whether PCA had affected the structure of the PV performance time series, the PCA-processed iPR^*_{PCA} was transformed back to $P^*_{A_{PCA}}$. The new $P^*_{A_{PCA}}$ time series were used to create monthly PR^*_{PCA} time series as per the formal definition of PR. The monthly time series were checked for the presence of trends through the univariate non-parametric Mann-Kendall test [170, 171] which is commonly employed to detect monotonic trends in series of environmental data, climate data or hydrological data.

The null hypothesis, H_0, is that the data come from a population with independent realizations and are identically distributed. The alternative hypothesis, H_A, is that the data follow a monotonic trend. The p-value of the Mann-Kendall statistic therefore represents the probability that the trend differs from zero. If the p-value is less than or equal to the chosen significance level ($\alpha = 0.05$), the test suggests that the observed data is inconsistent with the H_0, so the H_0 must be rejected. However, that does not prove that the tested hypothesis is true. When the p-value is calculated correctly, this test guarantees that the Type I error rate is at most α. For typical analyses, using the standard $\alpha = 0.05$ significance level, the H_0 is rejected when $p > 0.05$ and not rejected when $p < 0.05$.

Table 5.5 lists the resulting p-values of the test statistic. It can be concluded that by applying PCA, the original trends had disappeared for all systems except ucy12, ucy13 and ucy14. This essentially meant that the trend was captured in one of the lower PCA components and discarded as uncertainty. Unfortunately, due to the nature of PCA, the trend could not be pinpointed to a specific principal component. Therefore, more robust methods were employed.

Table 5.5: p-value from the univariate Mann-Kendall test for monotonic trend after PCA reconstruction.

System	PR	PR^*	PR^*_{PCA}
ucy04	0.0112	0.5120	0.7521
ucy05	0.0024	0.0010	0.2799
ucy06	0.1586	0.1716	0.4181
ucy07	0.0068	0.0051	0.1650
ucy08	0.0536	0.0650	0.4090
ucy09	0.1602	0.0158	0.4557
ucy10	0.0002	0.0000	0.1979
ucy11	0.0139	0.0079	0.3143
ucy12	0.0000	0.0000	0.0000
ucy13	0.0000	0.0000	0.0002
ucy14	0.0000	0.0000	0.0076

5.3.3 Robust principal component analysis

RPCA [91] is a modification of the PCA which works well with respect to grossly corrupted observations, such as in the case of PV performance metrics. RPCA decomposes a data matrix, $\mathbf{M}$, into a low-rank matrix, $\mathbf{L}$, plus a sparse matrix, $\mathbf{S}$, using the augmented Lagrange multipler (ALM) method [172].

Using RPCA, $\mathbf{M}$ was decomposed into $\mathbf{M} = \mathbf{L} + \mathbf{S}$, where $\mathbf{L}$ was a dense, low-rank matrix and $\mathbf{S}$ was a sparse matrix of perturbations. The use of RPCA in this case was intuitive, since the outliers, as well as the measurement points which were marked with a sentinel value, would be placed into the $\mathbf{S}$ matrix automatically by the procedure. They would then have the resulting influence on the inferred low dimensional subspace dropped.

$\mathbf{L}$ was recovered by the SVD, as in Eq. 5.7. The RPCA algorithm was applied by solving the convex programme called principal components pursuit (PCP) to minimize $\|L\|_* + \lambda\|S\|_1$, subject to $\mathbf{L} + \mathbf{S} = \mathbf{M}$, where $\lambda = 1/\sqrt{\max \dim \mathbf{M}} = 1/\sqrt{3287}$, as suggested in [91].

The application of RPCA on the wide matrix of each metric had proven to be extremely slow. Whereas classical PCA could decompose the matrix $\mathbf{M}$ in 10 ms, RPCA required 70 s and therefore could not be considered for this reason.

This is discussed in more detail in Appendix A.2.2.

5.3.4 Randomized robust principal component analysis

Randomized robust principal component analysis (rRPCA) is a method which uses the randomized inexact augmented Lagrange multiplier (IALM) method for obtaining the robust separation [173] in the same manner as the RPCA. These techniques exploit modern computational architectures more fully than classical methods [174] and open the possibility of dealing with truly massive data sets. In many cases, this approach provided better accuracy, robustness, and/or speed than its classical competitors, as randomized algorithms required $\mathcal{O}(mn\log(k))$ floating-point operations (flops) in contrast to $\mathcal{O}(mnk)$ for classical

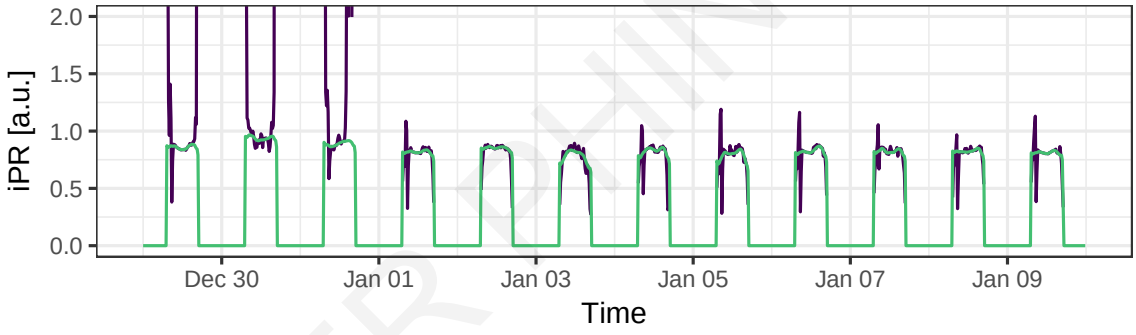
algorithms [175].

RRPCA was computed using a fast randomized algorithm (rsvd) to compute the approximate low-rank SVD decomposition. The sampling distribution of the randomized singular value decomposition (rSVD) was uniform in $[-1, 1]$.

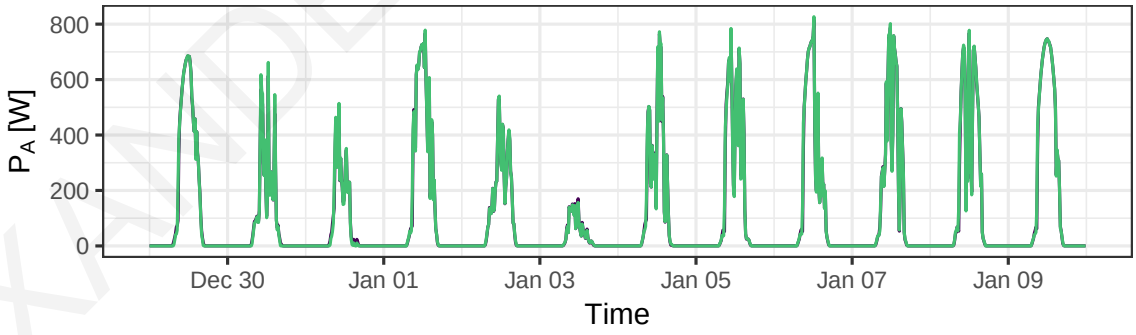
The results for all PV systems have shown that the originally rank 57 to 61 matrices of centred, outlier-corrected iPR^* were recovered by rank 23 to 26 $\mathbf{L}$ matrices and 62 % sparse, i.e. 62 % zero element, matrices $\mathbf{S}$.

Reconstruction from rRPCA

The fifteen minute iPR^*_{rRPCA} was reconstructed by rRPCA by subtracting the sparse perturbation matrix, $\mathbf{S}$, from the rectangular matrix, $\mathbf{M}$. The $P^*_{A_{rRPCA}}$ was then back-transformed from the iPR^*_{rRPCA}. The plots in Fig. 5.10 show a typical example of the uncertainty mitigated by the approach on ucy13. It can be seen that high variability on the iPR^* was effectively removed. These differences could more easily be assessed on the iPR^*_{rRPCA} than the back-transformed $P^*_{A_{rRPCA}}$.



(a) rRPCA on iPR^*



(b) Reconstruction of $P^*_{A_{rRPCA}}$

Figure 5.10: Results of applying rRPCA on ucy13 iPR^*.

The rRPCA can also work in an online fashion since it was designed to provide the fastest robust separation into principal components.

Suitability for degradation studies

In similar fashion to the PCA, the Mann-Kendall test was employed to check for presence of trends on the iPR^*_{rRPCA}. Table 5.6 lists the resulting p-values of the test statistic. It can

Table 5.6: p-value from the univariate Mann-Kendall test for monotonic trend after rRPCA reconstruction.

System	PR	PR^*	PR^*_{rRPCA}
ucy04	0.0112	0.5120	0.6729
ucy05	0.0024	0.0010	0.0013
ucy06	0.1586	0.1716	0.2131
ucy07	0.0068	0.0051	0.0076
ucy08	0.0536	0.0650	0.0868
ucy09	0.1602	0.0158	0.0045
ucy10	0.0002	0.0000	0.0000
ucy11	0.0139	0.0079	0.0096
ucy12	0.0000	0.0000	0.0000
ucy13	0.0000	0.0000	0.0000
ucy14	0.0000	0.0000	0.0000

be concluded that by applying rRPCA, the original trends were retained in the constructed monthly PR time series.

5.3.5 Comparison of the methods

A comparison of the boxplot and the cascade of boxplot + rRPCA on iPR is shown in Fig. 5.11 during a typical winter period and Fig. 5.12 during a typical fault period. The back-trasformation to P_A is shown in Fig. 5.13 and Fig. 5.14. From the figures, it can be concluded that including the boxplot outlier rule in the cascade resulted in being able to also estimate the period of faults and impute the expected performance for six out of the eleven PV systems in this study. The addition of rRPCA, resulted in mitigating the uncertainty of the fault detection.

5.3.6 Uncertainty

The uncertainty of applying the boxplot outlier rule on its own and the cascade of boxplot + rRPCA was evaluated through the standard deviation of the residuals. Fig. 5.15 plots the σ for the fifteen-minute residuals of each PV system's P_A and its back-transformation to P^*_A and $P^*_{A_{rRPCA}}$. As can be concluded, less amount of uncertainty was offered by the cascade Boxplot + rRPCA approach.

5.4 Missing data

5.4.1 Introduction

In the presence of missing data, the incomplete data sets can affect the R_{D_E} estimate, either by exacerbating the effect of the outage factors when the missing data point is of an instance of optimal operation, or by overestimating PV system performance and therefore

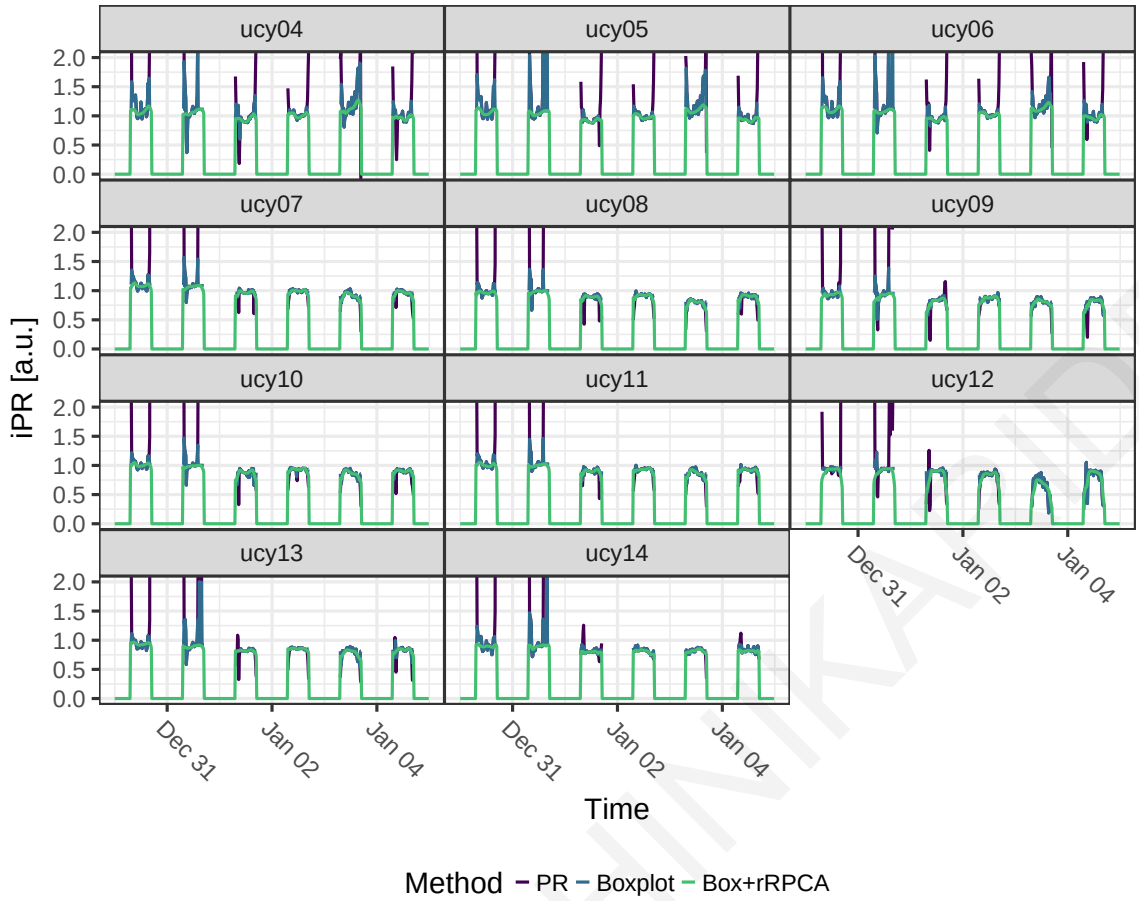


Figure 5.11: Performance of Boxplot and Boxplot+rRPCA on the *iPR* during a typical winter period.

underestimating degradation, when the missing data point is of a period of increased system losses. Therefore the effect of missing data on the underlying degradation will need to be investigated.

In this section, P_A time series were created with varying factors of missing data, which were randomly selected. Each time series was imputed with three different methods and the R_{D_E} was estimated on the permutation of the created time series.

5.4.2 Generation of artificially missing data

A Monte Carlo approach was employed, in order to create data sets with artificial missing data from the complete data set of observed values, A_{obs}, for each PV system. The index of the complete data set was randomly sampled at different levels in order to extract artificial outage periods. Random sampling without replacement was performed from 1 % to 40 % of the total amount of data in the complete data set. The data points, M, were then designated as missing from the complete data sets. The resulting incomplete data sets contained instances where all variates were missing at once (case deletion), to simulate multiple failures in the whole measurement chain. Since random sampling was utilized, an unbiased set of data points was assumed to be selected. This method represents the case where data

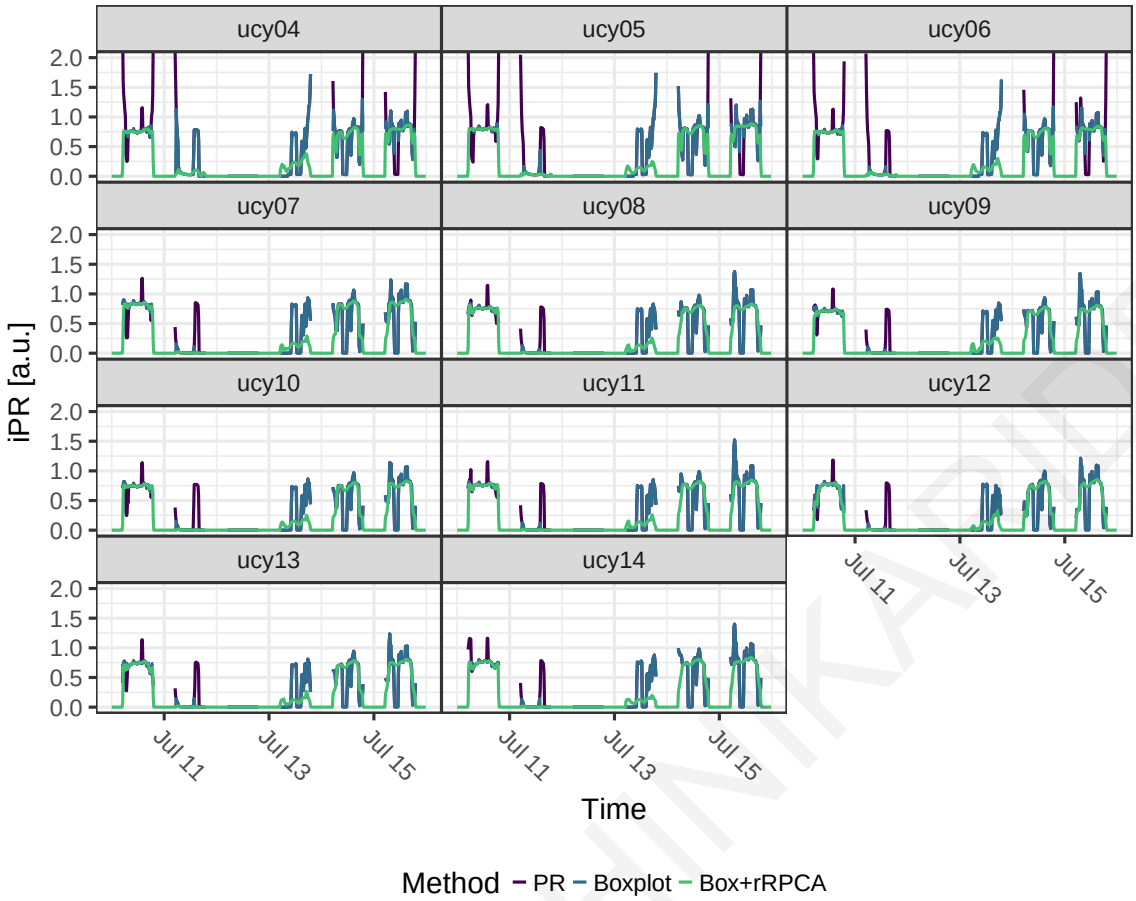


Figure 5.12: Performance of Boxplot and Boxplot+rRPCA on the *iPR* during a typical fault period.

is missing completely at random (MCAR) [94] and the distribution of missing data points did not depend on either the observed values or the missing values, as in Eq. 5.8.

$$P(M|A^{obs}) = P(M) \tag{5.8}$$

Lastly, as with the case of complete data sets, the incomplete data sets were used to create monthly *PR* time series for each simulated level of missing data. In total, forty incomplete data sets were created for each of the PV systems under study.

5.4.3 Imputation of missing data

As the estimation of the R_{D_E} relies on statistical analysis, the statistical properties of the time series should be retained without introducing bias by the imputation method. When using PV models such as PVUSA, single-point efficiency or others [176] to interpolate missing values, all the required explanatory variables may not always be available and may not always be measured (e.g. irradiance, module temperature, wind speed, humidity.) Additionally, since data logger, data transmission and storage related errors affect all measured variables, PV models were not applicable when a whole row of data was missing. Instead, the handling of missing data was based on univariate imputation of missing G_I and P_A data

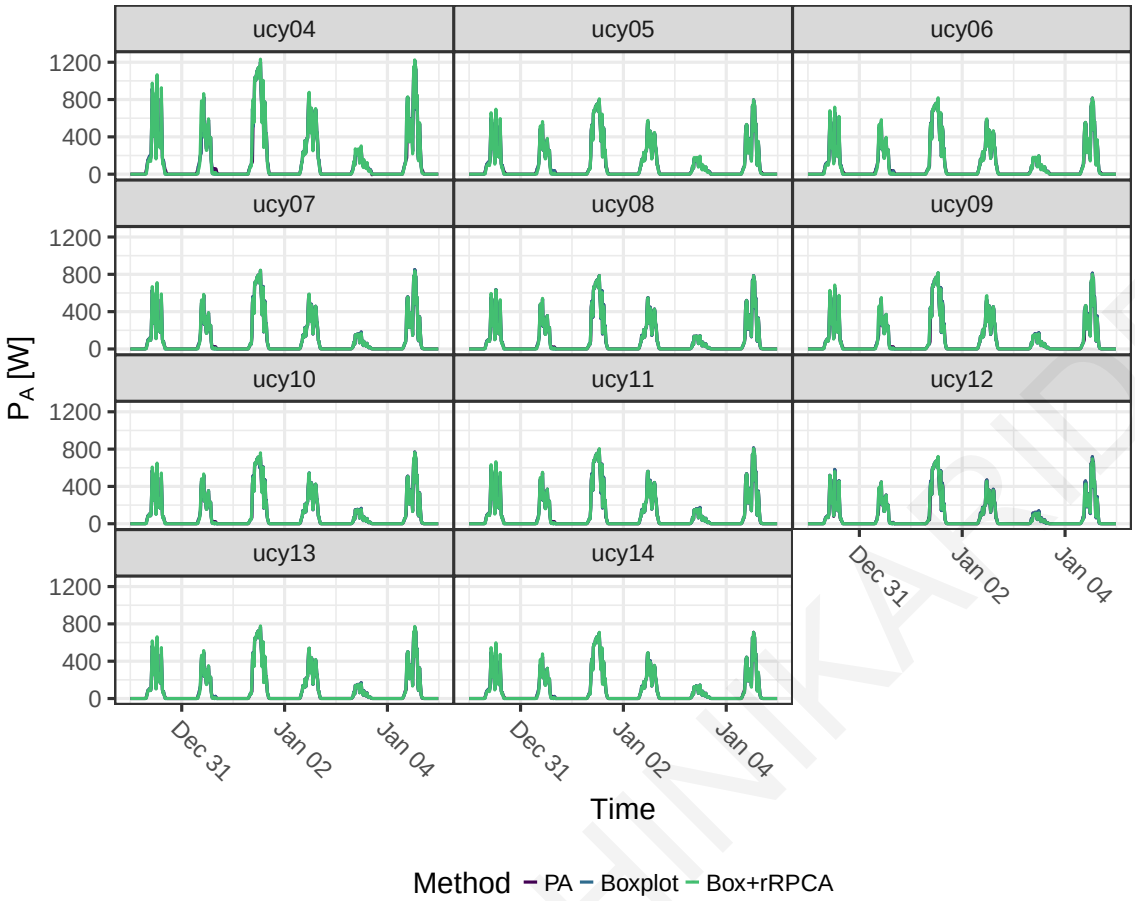


Figure 5.13: Performance of Boxplot and Boxplot+rRPCA on the P_A during a typical winter period.

points, where the basic assumption was that the pattern of missing data was independent of the underlying data set. The missing data points were imputed with a) the unconditional mean of the variable, b) LI, and c) bootstrapping in order to fill in the gaps in the fifteen-minute data sets. Firstly, the missing data points were imputed by the unconditional mean of the data set. This ensured that the mean of the measurement variable remained the same, but at the expense of large errors due to the seasonal profile of P_A and G_I. Similarly, the same missing data points were imputed by LI regressed on the time variable, in order to better model the effect of the trend present in the complete data set, as in Eq. 5.9.

$$P_A = \beta_1 t + \epsilon \tag{5.9}$$

where β_1 is a regression coefficient, t is the time and ϵ are the residuals.

Multiple imputations were created for each run, to estimate the uncertainty in the missing values.

5.4.4 Imputation by the bootstrap

In addition to imputation by the mean and linear interpolation, the bootstrap method was employed in order to use as much information from the distribution of each time series and

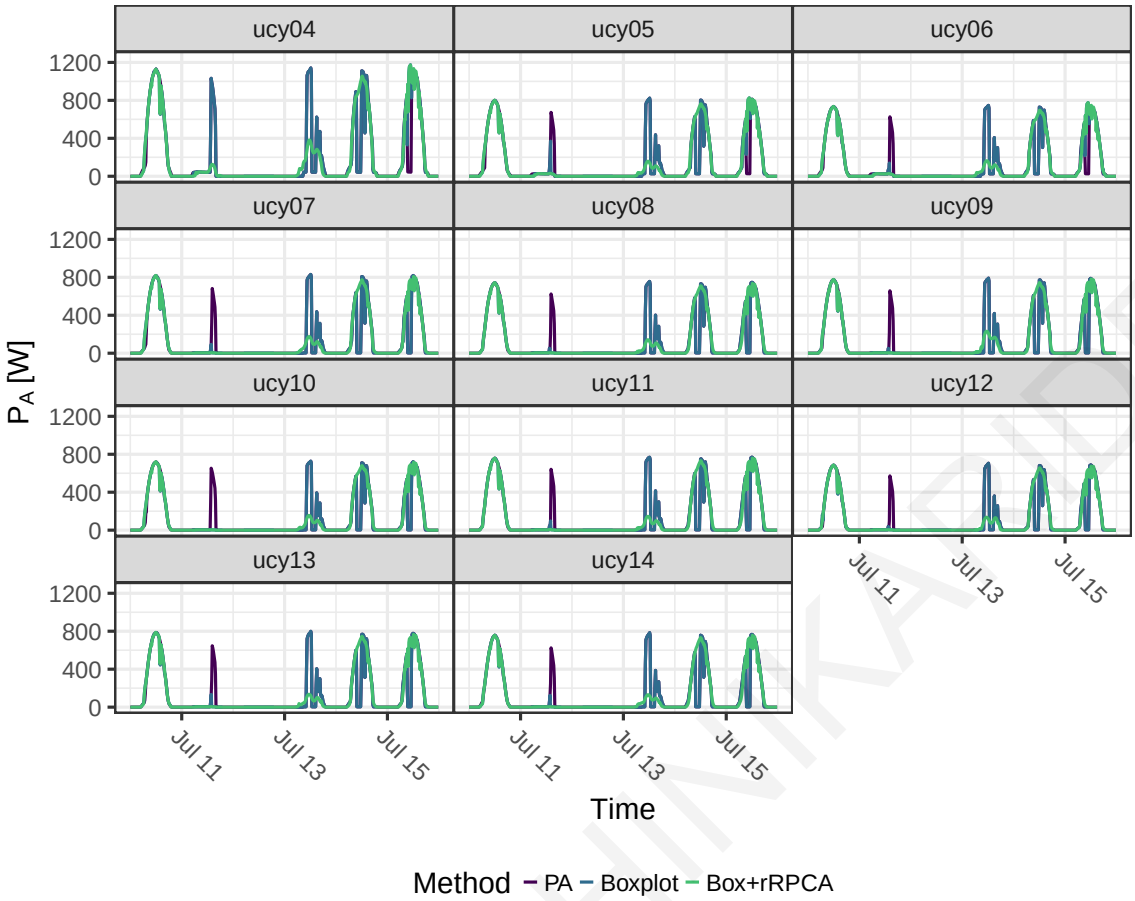


Figure 5.14: Performance of Boxplot and Boxplot+rRPCA on the P_A during a typical fault period.

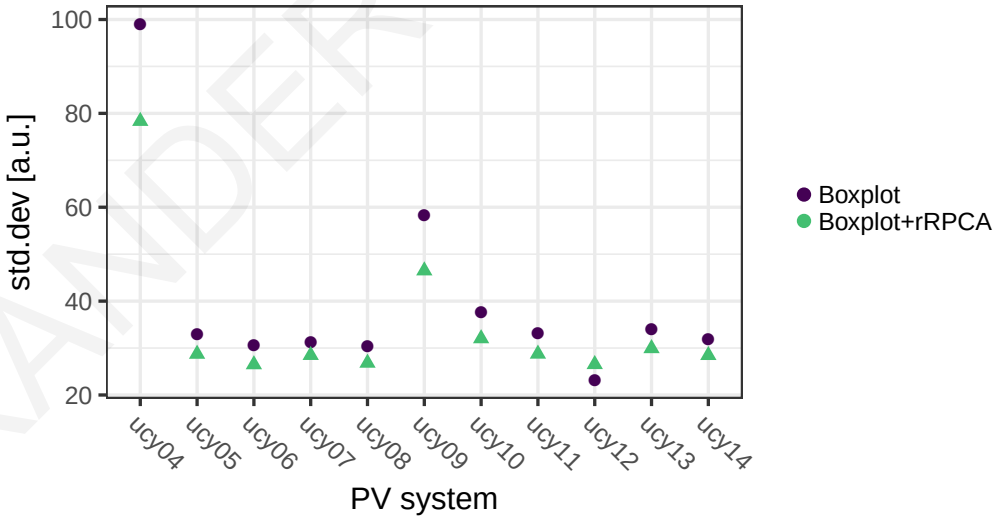


Figure 5.15: Standard deviation, σ, of the fifteen-minute residuals of each PV system's P_A^* and $P_{A_{rRPCA}}^*$.

be able to assess the uncertainty of the imputed values [98]. The bootstrap method relies on sampling from the posterior distributions and replacing missing data points with the sampled values [177]. The algorithm then re-evaluates and resamples the posterior distribution, and replaces the missing data points with the newly sampled values. This iterative

procedure is performed until the Expectation Maximization algorithm [178] converges. The result is that the missing values are filled in with a distribution of imputations that reflect the uncertainty about the missing data, for each data point.

The bootstrap method was applied on the data sets using the MICE package [179, 180] from *R* [181].

5.4.5 Effect of imputation on the degradation rate estimate

A large number of data sets were created by applying the methodology described in the previous subsection. For each PV system, 161 different data sets were analysed, in order to estimate the energy degradation rate: a) one complete data set, b) forty incomplete data sets with 1 % to 40 % of missing data, c) forty data sets imputed by the mean, d) forty data sets imputed by LI, and e) forty data sets imputed by the bootstrap.

Each data set of G_I and P_A values was used to create monthly PR time series which were then analysed with different statistical methods over the whole evaluation period. The PR time series for each PV array was seasonally adjusted with regression model with ARIMA errors (regARIMA) and CSD in order to separate it into the trend, the seasonal and the irregular components and was also modelled with LR using OLS.

The results of the analysis for 1 % to 40 % of sampled data points missing completely at random are shown in Fig. 5.16, Fig. 5.17 and Fig. 5.18 for the LR, CSD and regARIMA methods respectively. It can be concluded that imputation by the mean and LI performed very

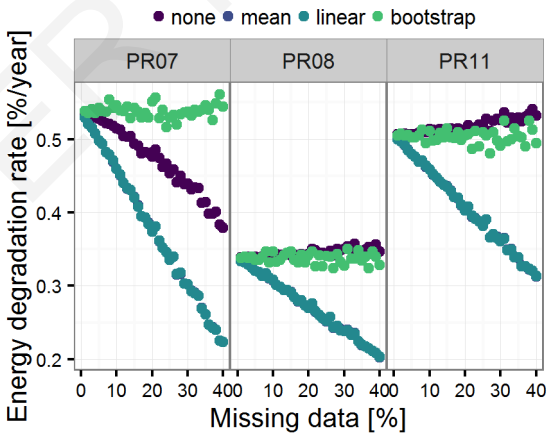


Figure 5.16: R_{D_E} estimated with linear regression for 1 % to 40 % missing data points and imputation by the mean, linear interpolation and bootstrap.

poorly for all PV systems as these two methods underestimated the R_{D_E} consistently with increasing amount of missing data. On the other hand, imputation by the bootstrap has been shown to provide an improvement to the estimation of R_{D_E}, across all three degradation estimation methods. For LR, the bootstrap provided robustness for up to 20 % of missing data for the ucy07 and ucy11 systems, whereas for the ucy08 system, the estimates were stable even at 40 % of missing data.

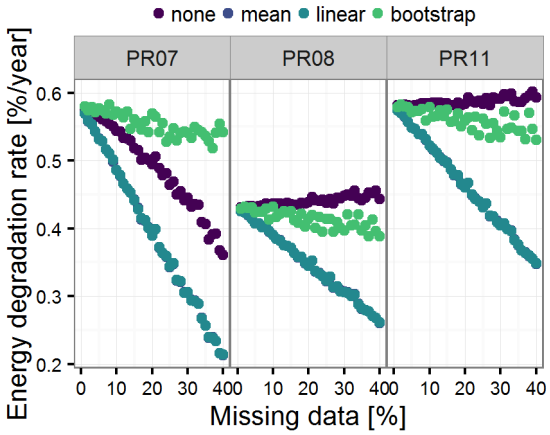


Figure 5.17: R_{D_E} estimated with CSD for 1 % to 40 % missing data points and imputation by the mean, linear interpolation and bootstrap.

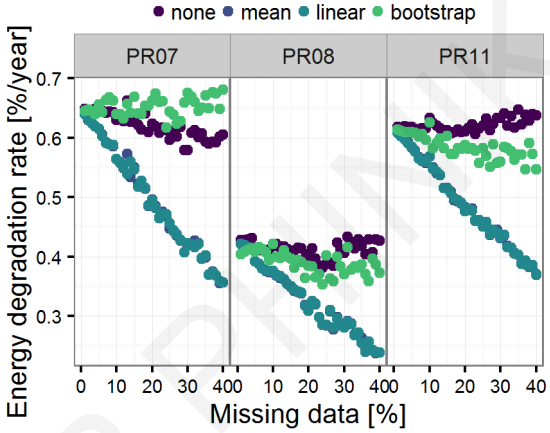


Figure 5.18: R_{D_E} estimated with regARIMA for 1 % to 40 % missing data points and imputation by the mean, linear interpolation and bootstrap.

The results of applying CSD were very sensitive to the amount of missing data and the application of imputation. Without imputation, the estimates were underestimated by 0.22 %/y for the ucy07 system and overestimated by 0.05 %/y for the ucy08 and ucy11 systems. By using the bootstrap, the under/overestimation was reduced to 0.05 %/y for all systems. In the case of regARIMA, the results were shown to be the most robust to missing data for all systems, regardless of whether imputation was used. For the ucy07 and ucy11 systems, the maximum differences without imputation were 0.05 %/y and for the ucy08 the maximum difference was 0.02 %/y. When imputed by the bootstrap, the differences were comparable but lower.

A comparison between degradation estimation methods with incomplete data sets, where no imputation was applied, has shown that regARIMA was the most robust, for up to 10 % of missing data for the ucy07 system and 20 % of missing data for the rest of the systems. LR and CSD were shown to be very sensitive to missing data, starting from 2 %.

Finally, to mitigate the effects of missing data, imputation by the bootstrap was shown to be the most successful in providing robust energy degradation rate estimates. RegARIMA

was the most robust method, regardless of whether imputation was used. LR was more robust when missing values were imputed by the bootstrap and CSD was the most sensitive method to missing data.

As the results have shown, this method outclassed conventional linear interpolation, historical average, and industry best practice imputation approaches in dealing with missing data and imputing PV power measurements. The end conclusion was that when using regARIMA up to 10 % of data could be missing, before having to apply any imputation on the dataset.

5.5 Conclusions

In this chapter, a simple and applicable methodology has been developed which treats the problem of outlier detection as one-class classification problem. Three statistical methods have been tested for applicability in the field of PV and used to develop the final methodology, taking into account the density of measurement data, minimization of the number of covariates, the effect of missing data, the simplicity of the approach and its fast convergence for online applicability.

The flowchart of the developed methodology is shown in Fig. 5.19. This proposed flowchart could easily be implemented in an online fashion.

Finally, through the work developed in this chapter, the need for ad hoc corrections on the energy yield was eliminated, in favour of more general procedures which can also provide inference.

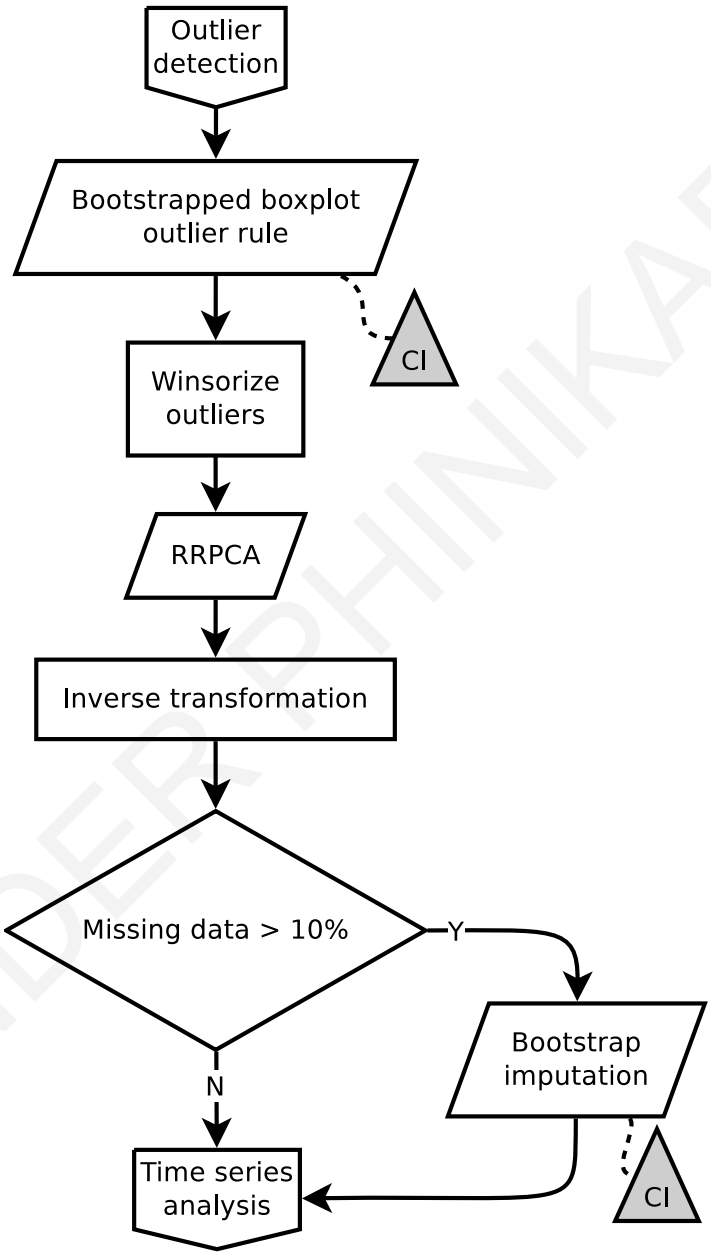


Figure 5.19: Flowchart of the developed outlier detection methodology.

Chapter 6

Time Series Analysis

Work from this chapter has been published in [182, 183, 184, 70, 103, 185]

6.1 Introduction

In order to estimate degradation through field measurement data, the Energy Degradation Rate, R_{D_E}, quantity was defined. The R_{D_E} characterizes the degradation of the field performance, estimated through analysis of measurements from a PV array/system under a broad spectrum of prevailing meteorological conditions.

> **Definition (Energy degradation rate):** A scalar quantity which is defined as the annual, linear percentage reduction of the field performance metric.

In this work, the performance metric was the monthly *PR*, as formally defined in IEC 61724:1998 [36], using fifteen-minute average data of G_I and P_A or $P^*_{A_{rRPCA}}$ as defined in Sec. 5.3.4.

Due to the prevailing seasonal profile of PV power production, the calculation of R_{D_E} was generalized for multiple geographical regions by designing proper statistical analyses to decouple the seasonal component from the underlying trend of PV performance. Such statistical methods were initially developed to test hypotheses about the data and were the basis for econometrics and other time-series heavy disciplines. The general structure of PV field performance data sets is closely related to economic time series, with very similar characteristics, making the application of similar time series analysis methods a reasonable option, although in PV there is the advantage of knowing a-priori which independent variables could explain the data. Therefore, combining the physical models of PV performance [186] with statistical data analysis methods allows additional information to be extracted from time series of PV performance in the field, as was shown by the definition of the *iPR* in this work.

In this chapter, the viability of statistical models for estimating a trend on monthly *PR* time series is presented. The modelled trends are assessed for compliance to the underlying

assumptions and tested for non-linearity and changepoints. Finally, the uncertainty of each method is estimated by bootstrap.

6.1.1 Indication of trend

Prior to performing any R_{D_E} estimation, the time series is checked for the presence of trends via the Mann-Kendall test, described in Sec. 5.3.2. As with all non-parametric tests, the Mann-Kendall test does not assume a specific distribution for the population. The R_{D_E} was estimated through the parametric and non-parametric models described in this chapter and then its confidence was estimated by the Mann-Kendall test statistic.

In Table 5.6 it has been shown that the p-value of the test statistic was greater than the 0.05 significance level for ucy04, ucy06 and ucy08, which failed to reject the H_0 and suggested that the PR values of those systems were independent and identically distributed (i.i.d.) (i.e. no trend could be detected with confidence.)

When the given time series is of a monthly seasonal structure, the H_0 given above may be too restrictive and fail to detect any trend. The seasonal Mann-Kendall test [187, 188] was employed to test for seasonal trends, where the Mann-Kendall statistics are computed for each season separately. For these time series, season refers to each of the twelve months of the year. The p-values of the seasonal Mann-Kendall test statistic had shown that only the ucy10, ucy12, ucy13 and ucy14 PR and PR^*_{rRPCA} time series featured monotonic negative seasonal trends.

In the presence of seasonal correlation, neither the seasonal nor the non-seasonal Mann-Kendall test could be considered an exact test. The correlated seasonal Mann-Kendall test was thus employed, since in the case of PR time series, autocorrelation was detected in the data via the ACF. This test statistic suggested the presence of trend on all PR and PR^*_{rRPCA} time series, except ucy06 and ucy09.

6.2 Methods for estimating the degradation rate

The following methods were employed in an attempt to model the trend of the PR and PR^*_{rRPCA} time series:

- Linear regression (LR)
- Classical Seasonal Decomposition (CSD)
- Seasonal-Trend Decomposition by LOESS (STL)
- Theil-Sen estimator (TS)
- X-13ARIMA-SEATS (regARIMA)

6.2.1 Non-parametric methods

Non-parametric filtering methods are inherently different than stochastic model-based methods because an explicit model is not specified. Such a method is seasonal-trend decomposition by LOESS (STL) [189], which extracts the trend from locally weighted polynomial fitting [190, 191]. STL can provide robust estimates of the trend and seasonal components that are not distorted by outliers and missing values. By using the STL [189], a time series can be decomposed into seasonal, trend and irregular components. This iterative filtering procedure estimates the seasonal component by smoothing the seasonal sub–series (e.g. the series of all solar noon values). In each iteration of the procedure, the moving average smoothing and LOESS smoothing are used multiple times. In the case of STL, the R_{D_E} is estimated as the negative slope of the trend component.

The Theil-Sen estimator (which is referred to as the Y-o-Y method in PV literature and described in Ch. 2) is also another non-parametric method which is robust to outliers since it chooses the median slope among all lines passing through the data points. The slopes of all lines are calculated as follows:

$$d_k = \frac{PR_j - PR_i}{j - i} \tag{6.1}$$

for $(1 \leq i < j \leq n)$, where d is the slope, PR is the monthly Performance Ratio, n is the number of points and i and j are indices. The R_{D_E} is thus equal to the TS slope which is the median from all slopes, multiplied by 12 to convert to an annual value and by 100 to convert to percentage:

$$R_{D_E} = -1 * 12 * 100 * \text{Median}(d_k) \tag{6.2}$$

6.2.2 Linear Regression

The R_D method is the most widely used in the literature, as detailed in Ch. 2. It was applied through OLS on monthly PR time series:

$$\hat{PR} = \beta_1 t + \beta_2 + \epsilon \tag{6.3}$$

where $\hat{PR}$ represents the fitted monthly PR, β_1 is the slope of the trend and β_2 is the intercept. The R_{D_E} was thus estimated by:

$$R_{D_E} = -1 * 12 * 100 * \beta_1 \tag{6.4}$$

6.2.3 Classical Seasonal Decomposition

The CSD method is based on the application of a centred moving average and calculation of seasonal indices by averaging the extracted seasonal component for each month. It assumes that the seasonal component of PV performance remains stable year after year. An additive

model was specified, as the annual seasonal component due to prevailing weather was expected to be relatively stable:

$$\hat{PR} = T_t + S_t + e_t \tag{6.5}$$

where $\hat{PR}$ is the fitted monthly PR, T_t represents the trend, S_t represents the seasonal component and e_t the residual component. The trend was calculated by applying a 12-month centred MA to monthly PR time series. For a $2 \times k$ moving average, where k is the seasonal period ($k = 12$ because of the number of months in a year), the centred MA at time t was calculated by:

$$T_t = \frac{1}{2}\left(\frac{1}{k}\sum_{i=t-m}^{t+m-1} PR_i + \frac{1}{k}\sum_{i=t-m+1}^{t+m} PR_i\right) \tag{6.6}$$

where T_t is the trend at time t, ($t > m$), and m is defined as the half-width of a moving average, $m = k/2$. The R_{D_E} was thus estimated as the linear slope of the trend component, multiplied by 12 to convert to an annual value and by 100 to convert to percentage:

$$\hat{T}_t = \beta_1 t + \beta_2 + \epsilon \tag{6.7}$$

$$R_{D_E} = -1 * 12 * 100 * \beta_1 \tag{6.8}$$

6.2.4 Autoregressive integrated moving average models

ARIMA models were popularised by Box and Jenkins [113]. An ARIMA model describes a univariate time series as a combination of AR and MA lags which capture the autocorrelation within the time series. The order of integration denotes how many times the series has to be differenced to obtain a stationary series. The augmented Dickey–Fuller (ADF) [192] test is used to test for a unit root in the series and determine the order of differencing. Another stationarity test, the Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test [193], uses the H_0 that the series is trend stationary (stationary around a deterministic trend). By using both the ADF and the KPSS, series that appear to be stationary, series that appear to have a unit root, and series for which it cannot be ascertained whether they are stationary or integrated can be distinguished. For the PV systems used in this work, both tests had shown that there were unit roots and deterministic trends, as can be seen in Table 6.1 which lists the order of differencing determined using each test.

ARIMA models are associated with the Box-Jenkins approach to time series. According to this approach, the series must be differenced until stationary, and then the ACF and PACF plots of the stationary series should be consulted to determine the appropriate lag order for an ARIMA process. The model parameters are then estimated, and the model must be checked for autocorrelation in the residuals. As explained before, the ARIMA model has free parameters which have to be selected by the designer of the system manually with the Box-Jenkins approach. To automate the approach, various systematic techniques

Table 6.1: Number of differencing orders required, as determined by ADF and KPSS tests.

System	PR (ADF)	PR (KPSS)	PR^{*}_{rRPCA} (ADF)	PR^{*}_{rRPCA} (KPSS)
ucy04	1	0	1	0
ucy05	1	0	1	1
ucy06	1	0	1	0
ucy07	1	0	1	0
ucy08	0	0	0	0
ucy09	0	0	0	0
ucy10	0	0	0	1
ucy11	0	0	0	0
ucy12	0	1	0	1
ucy13	0	1	0	1
ucy14	0	1	0	1

can be utilised. In this work, a regression model with ARIMA errors, or regARIMA, using the seasonal adjustment technique developed by the U.S. Census Bureau (X-12-ARIMA) which was used to estimate a non-linear, non-monotonic trend from time-series data, while removing all autocorrelation in the residuals and automatically determining the order of the parameters via the AICc.

A regARIMA model estimates the mean of the time series, Y_t, as a linear combination of regressors and the residual component, z_t, by an ARIMA process, as in Eq. 6.9:

$$Y_t = \sum_i \beta_i x_{i_t} + z_t \tag{6.9}$$

where Y_t is the monthly PR time series and the residual component, z_t, is modelled by the seasonal (multiplicative) ARIMA model given in Eq. 6.10:

$$\Phi(T)\Phi_S(T_S)\nabla^d\nabla_S^D z_t = \theta(T)\theta_S(T_S)e_t \tag{6.10}$$

In this way, the regARIMA model residuals possess, by definition, Gaussian white noise (GWN) properties which therefore satisfies the most basic requirement of stochastic models.

Optimal regARIMA models were fitted to the data through the X-13ARIMA-SEATS algorithm [194] which was developed by the U.S. Census Bureau and used extensively in econometrics and time series analysis by institutions such as the Deutsche Bundesbank [195], the European Central Bank [196], the Australian Bureau of Statistics, Statistics Canada, Office of National Statistics UK and many others. The first seasonal adjustment method where regARIMA was derived from, was released in 1967 as the X-11 procedure [197], enhanced in 1983 with ARIMA forecasts and backcasts by the X-11-ARIMA procedure [198], further enhanced in 1998 by adding regARIMA, outlier identification and a plethora of diagnostics into X-12-ARIMA [199] and most recently in 2012 by the update of X-12-ARIMA with SEATS-TRAMO (X-13ARIMA-SEATS) [194, 200] method. X-11-ARIMA was developed to use ARIMA model forecasts to extend the original series at both ends, which is then sea-

sonally adjusted. Doing this mitigates the uncertainty in the trend estimation and seasonal averaging near the tails of the time series. These extensions were a very important improvement offered by X-13ARIMA-SEATS over CSD. Another advantage of X-13ARIMA-SEATS was that it was able to yield trend estimates for all observations, in contrast to CSD, which was unable to produce trend estimates for half of the seasonal period at each end of the time series.

The X-13ARIMA-SEATS algorithm behaves in a similar way to CSD, and then the components are refined through several iterations. The X-13ARIMA-SEATS modelling method allows the calculation of the optimal regARIMA model for each PV system time series and uses robust statistics to decompose the time series into the seasonal, the trend and the residual components. The following outline of the method describes an additive decomposition applied to monthly data, such as PR time series.

1. Test the stationarity of the time series with the ADF test and the KPSS test and transform it to achieve stationarity.
2. Detect and remove extreme values one-by-one [201, 202]
3. Iteratively fit an ARIMA process with varying model orders until the AICc is minimized
4. Subtract the residuals of the best fitted ARIMA model and calculate a 3x3 or 3x5 Henderson MA
5. Cross-validate the model by testing residual autocorrelation

In this way, the R_{D_E} estimated with regARIMA was calculated from the slope of the MA filtered trend, multiplied by 12 to convert it to an annual value and by 100 to convert to percentage.

Model selection

The X-13ARIMA-SEATS algorithm was applied in batch mode. Optimal model orders, which were estimated automatically via minimizing the AICc, are listed in Table 6.2 for each PV system PR and outlier corrected PR^*_{rRPCA}. The AICc penalizes high model orders, therefore the selected optimal models satisfied the principle of parsimony. The monthly time series have shown similar temporal characteristics between them, with many systems being optimally modeled with the same model structure.

The models were validated through the X-13ARIMA-SEATS diagnostics and by examining the properties of the residuals and the statistical significance of the model parameters. The Ljung-Box test was used to automate testing for autocorrelation in the residuals and determining which autocorrelation coefficients were significant in an unsupervised way. This provided a level of confidence for the validity of the model and the fit. Besides the Ljung-Box test, this procedure can be performed manually by constructing a plot of the

Table 6.2: Optimal regARIMA model orders.

System	PR	PR^{*}_{rRPCA}
ucy04	(1 0 0)(0 1 1)	(1 0 0)(0 1 1)
ucy05	(0 1 1)(0 1 1)	(1 0 1)(1 1 0)
ucy06	(0 1 1)(0 1 1)	(1 0 0)(0 1 1)
ucy07	(0 1 1)(0 1 1)	(1 0 0)(0 1 1)
ucy08	(0 1 1)(0 1 1)	(0 1 1)(0 1 1)
ucy09	(0 1 2)(0 1 1)	(0 1 1)(0 1 1)
ucy10	(1 0 1)(1 1 1)	(0 1 1)(0 1 1)
ucy11	(0 1 1)(0 1 1)	(0 1 1)(0 1 1)
ucy12	(0 1 1)(0 1 1)	(0 1 1)(0 1 1)
ucy13	(0 1 1)(0 1 1)	(0 0 0)(0 1 1)
ucy14	(0 1 1)(0 1 1)	(0 1 1)(0 1 1)

ACF of the residuals and observing at which lags there was significant autocorrelation. In this case, autocorrelation of the residuals was insignificant for all PV systems under study.

6.3 Uncertainty

Some of the effects of model uncertainty are too narrow prediction intervals, and the non-trivial biases in parameter estimates which can follow data-based modelling. One way to overcome the effects of model uncertainty would be to use the bootstrap.

6.3.1 Standard errors

From the modelled trends, the linear R_{D_E} was calculated from the linear slope of the trend. As such, the uncertainty of the slope was approximated using the standard error (SE) of the slope estimator of LR on the trend of each PV system time series [128] as in Eq. 6.11:

$$SE = \sqrt{\frac{1}{n-2} \times \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}} \tag{6.11}$$

where n is the sample size, y is the observed value and $\hat{y}$ is the estimated value of the dependent variable respectively, x is the observed value and $\bar{x}$ is the mean value of the independent variable respectively. The expanded uncertainty, U_{R_D}, can be calculated for a 95 % confidence level:

$$U_{R_D} = k \times SE \tag{6.12}$$

where $k = 2$, if the linear model assumptions are correct.

The interval created by $R_{DE} \pm 2 * SE$ is an approximation of the 95 % confidence interval. The coefficient standard errors, estimated by the model, rely on asymptotic approximations and may not be trustworthy in a sample of size 108 months, i.e. the length of the monthly PR and PR^{*}_{rRPCA} time series used in this work.

6.3.2 Bootstrap Confidence Intervals

Bootstrapping has been described in Sec. 5.3.1, where it was applied on boxplot statistics, to get a better estimate of the outlier thresholds and the confidence interval. In the context of modelling, bootstrapping can either be performed on the residuals of the regression models used in this analysis, namely the LR, CSD and X-13ARIMA-SEATS or on the time index for the STL. Regarding the TS, it was robust to the presence of outliers and additionally, since the slope was not a point estimate but a distribution of estimates, the confidence intervals reported by the method were considered unbiased.

When the linear model of Eq. 6.13 is used, the most intuitive way to construct a confidence interval, would be to use the residual bootstrap to model variation of the error term.

$$y_i = \mathbf{X}_i\boldsymbol{\beta} + e_i \tag{6.13}$$

Thus the bootstrapped linear model would become:

$$y_i^* = \mathbf{X}_i\hat{\boldsymbol{\beta}} + e_i^* \tag{6.14}$$

which assumes that the actual error terms are i.i.d., which is a strong assumption to make with the LR, CSD and STL models, as their Ljung-Box statistic and the ACF and PACF plots signified the presence of autocorrelation. On the other hand, if the model was correctly specified and designed to capture all the information in the residuals, then residual resampling would be a good option.

One alternative in the case of non-i.i.d. residuals would be to resample the time index, also called random-x-resampling. This was less intuitive because the time between successive measurements was fixed, so resampling in this way would construct an irregular time grid. Random-x resampling was performed on LR, CSD and STL models, since it has been shown that the model residuals had significant autocorrelation.

Finally, the confidence interval was created as in Eq. 5.6.

6.4 Trend estimation

6.4.1 Linear trend

In the context of comparing the estimated R_{D_E} to the linear warranties given by PV module manufacturers, a linear trend could be assumed. A linear trend assumes that the performance metric of a PV module/array/system drifts monotonically to a critical value, after which failure occurs. In this work, trend estimation was treated as a post seasonal-adjustment step, therefore a linear fit was applied on extracted trends which were not completely linear. These were the STL, CSD and X-13ARIMA-SEATS estimated trends. On the other hand, fitting the trend with a straight line had the advantage of being able to compare against experimental results which will be presented in Ch. 7.

The linear R_{D_E} is thus given in units of %/y and is defined as such:

$$R_{D_E} = -1 * 12 * 100 * \beta \tag{6.15}$$

where β is the slope coefficient of the trends extracted with LR, CSD. STL and X-13ARIMA-SEATS or the median of all slope coefficients estimated with TS, 12 converts the monthly slope to an annual value and 100 converts it to a percentage.

The linear R_{D_E} and its confidence interval for the LR, CSD, STL, X-13ARIMA-SEATS and TS methods is plotted in Fig. 6.1. The confidence interval shown was estimated by bootstrapped slope estimates through residual resampling for X-13ARIMA-SEATS and random-x-resampling for LR, CSD and STL. The CI for TS was provided by the CI of the estimator. From this plot, it can be seen that all three metrics, namely the PR, the Manually corrected PR (as described in Sec. 4.1) and the PR^*_{rRPCA} were able to provide similar estimates for every system under evaluation, except ucy04, and across all methods. The differences observed for the ucy04 system could be traced back to the Mann-Kendall test results in Table 5.6, which has shown that a trend could not be detected with confidence for ucy04 and ucy06.

From this plot, it is evident that the R_{D_E} of c-Si arrays hovered around the 0.7 %/y rate, with the exception of ucy10. On the contrary, higher R_{D_E} was estimated for all the thin-film technologies, with CIGS degrading between 2.28 %/y to 2.55 %/y, CdTe degrading between 1.48 %/y to 1.98 %/y and a-Si degrading between 1.11 %/y to 1.24 %/y

6.4.2 Non-linear trend

As described in Ch. 2, it is sometimes observed in PV performance time series, that the trend is not always a straight line. To assess the linear degradation assumption throughout the 9 y of operation of the systems, a non-linear trend estimation procedure was developed. A non-linear trend assumes that at some point in time, a change in the operational characteristcs, forced the trend in another direction. This point in time is called a change point and several statistical methods exist to estimate its position.

To detect a single change point in the trends, the non-parametric Pettitt's test was applied on the estimated trends [203] to test for a shift in the central tendency of the time series. This non-parametric test hypothesizes the H_0 that there is no change against the H_1 that a change point exists. In this way, the proposed method made it suitable to be employed in the unsupervised, online fashion.

Table 6.3 shows the change points and their significance, detected on the regARIMA trends from the PR^*_{rRPCA} time series of the systems under study. According to their p-values, these change points were all significant. These change points were used to segment the time series at the detected points and compute a separate R_{D_E} for each segment with the robust TS estimator, again, in an automated way.

The plots in Fig. 6.2 show the differences in slopes before and after the changepoint. Two

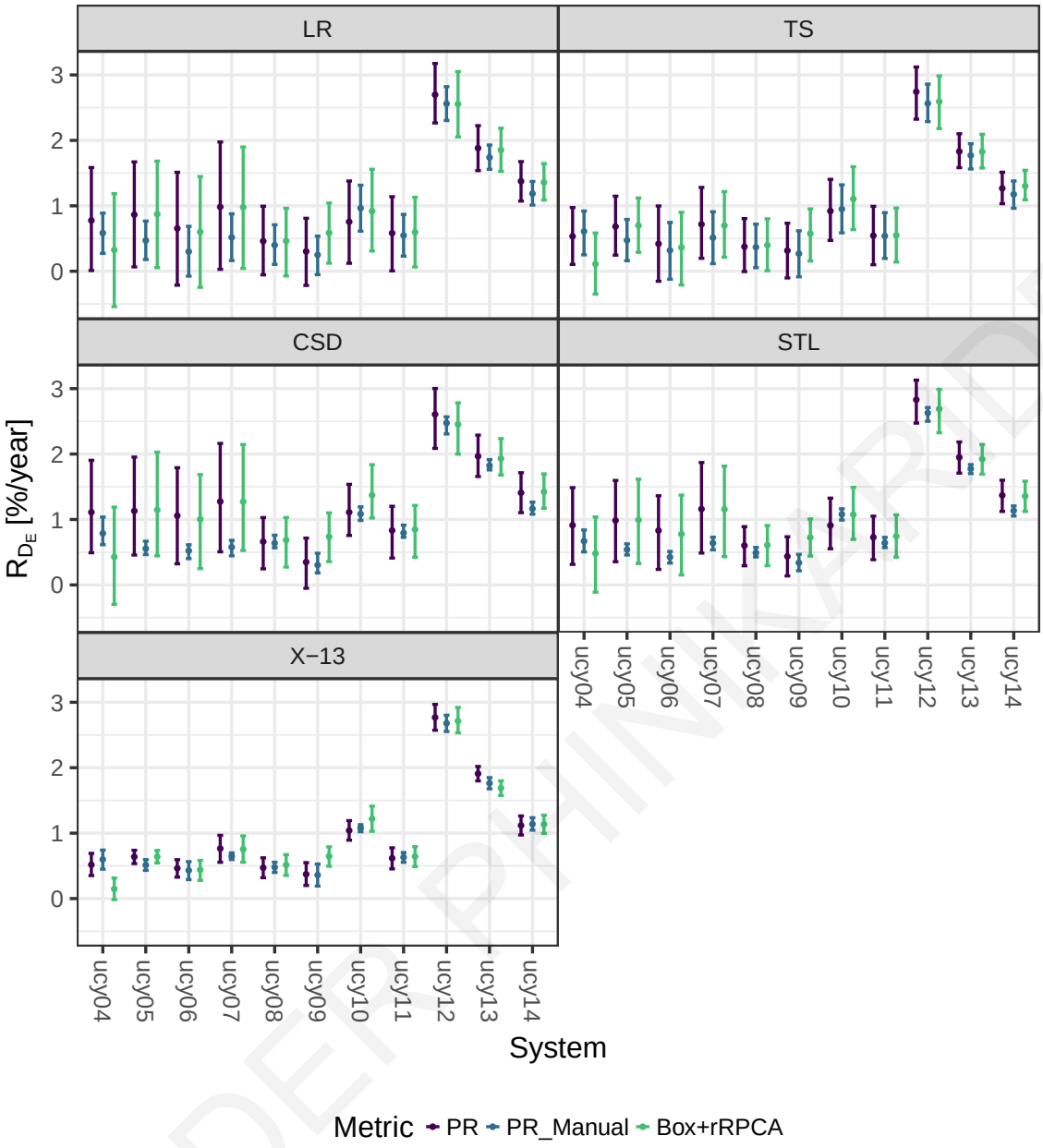


Figure 6.1: Linear R_{D_E} and confidence intervals, using LR, CSD, STL, TS and X-13ARIMA-SEATS on metrics of PR, Manually corrected PR and PR^{*}_{rRPCA}.

different degradation modes could be explained by the outdoor exposure time required for a PV module to reach a stable performance and by the effect of faults such as hotspots and cracked cells, as in the case of ucy09 and ucy10, which will be explained further in Ch. 7.

The estimated segmented degradation rates are listed in Table 6.4. From the results listed in this table and the slopes shown in Fig. 6.2, the importance of using a robust estimator for the slope was demonstrated.

6.5 Conclusions

This chapter has described the unsupervised and generalized methodology for assessing the R_{D_E} of fielded PV systems which was developed in this work. The methodology can

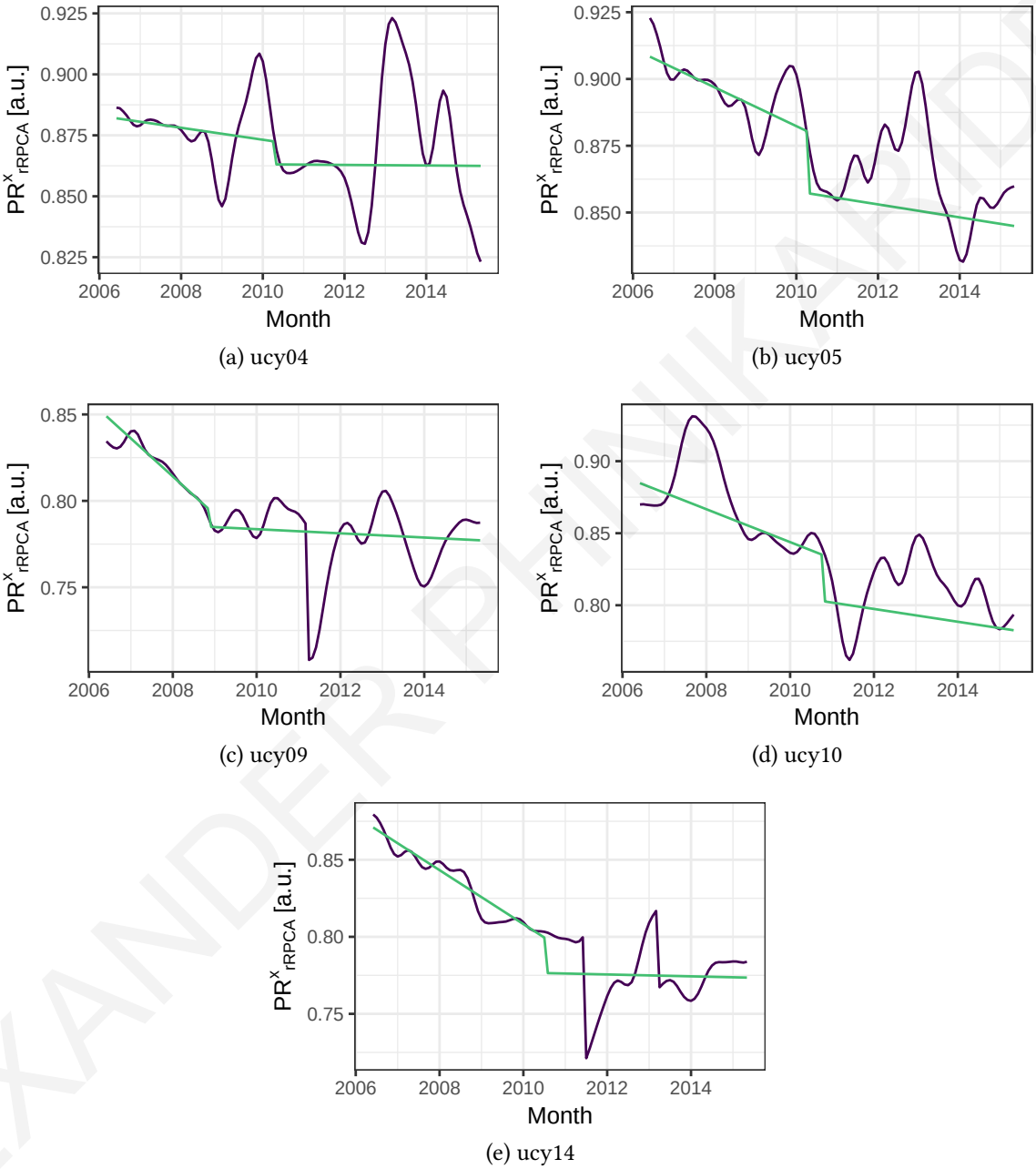
Table 6.3: Change points detected with the Pettitt test on the PR^{*}_{rRPCA} regARIMA trend.

System	Point in Time	p-value
ucy04	47	0.0068
ucy05	47	0.0000
ucy06	48	0.0000
ucy07	47	0.0000
ucy08	52	0.0000
ucy09	30	0.0000
ucy10	53	0.0000
ucy11	51	0.0000
ucy12	54	0.0000
ucy13	54	0.0000
ucy14	50	0.0000

Table 6.4: R_{D_E} before and after the change points detected with the Pettitt test on the PR^{*}_{rRPCA} regARIMA trend.

System	R_{D_E} before [%/y]	R_{D_E} after [%/y]
ucy04	0.2457	0.0119
ucy05	0.7255	0.2429
ucy06	-0.4324	-0.5760
ucy07	0.3851	0.5806
ucy08	0.5483	-0.4926
ucy09	2.1991	0.1208
ucy10	1.1426	0.4438
ucy11	0.5450	-0.3130
ucy12	1.9962	2.8654
ucy13	2.4224	0.6849
ucy14	1.7494	0.0611

be summarized with the flowchart shown in Fig. 6.3. This developed methodology can be completely automated, it can provide statistical inference and relies on non-parametric statistics which abstract the underlying distribution of data. Therefore, the proposed approach solves many of the problems found in the current literature and addresses many of the challenges of the PV degradation rate estimation field.



(a) ucy04

(b) ucy05

(c) ucy09

(d) ucy10

(e) ucy14

Figure 6.2: Trends estimated through X-13ARIMA-SEATS and segmented degradation slopes, based on Pettitt test change points.

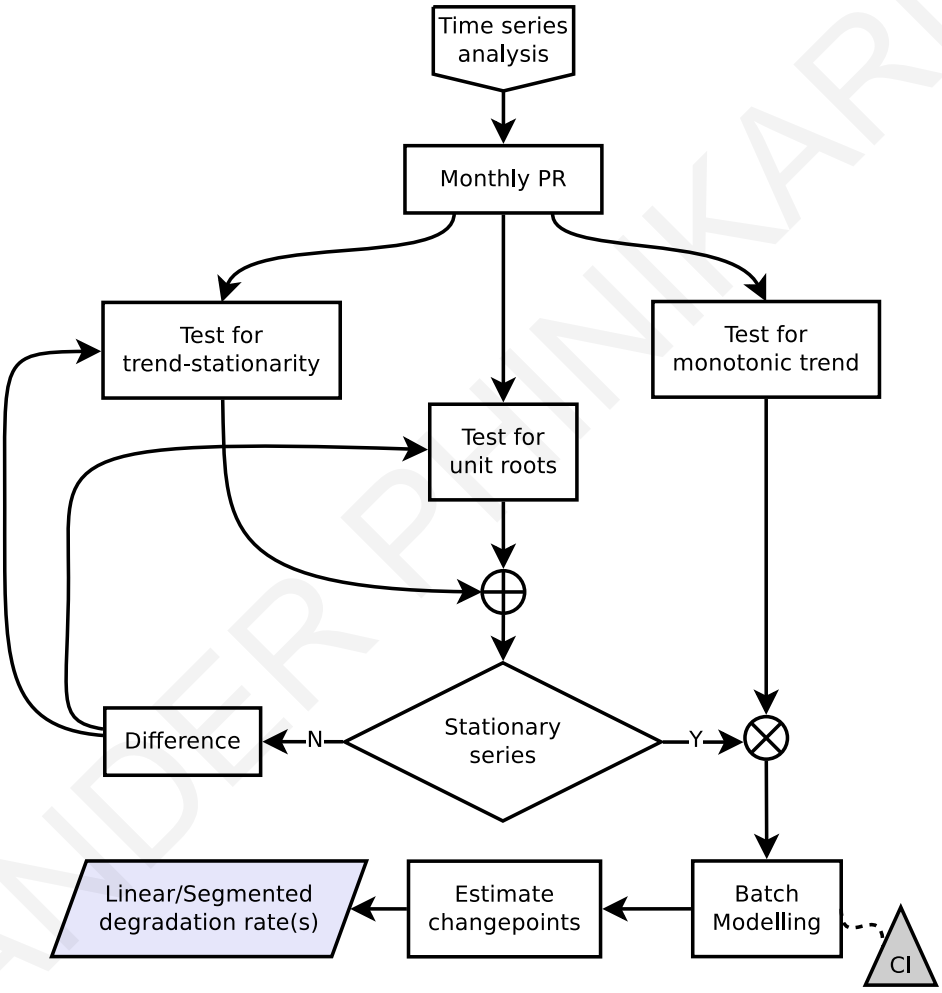


Figure 6.3: Flowchart of the developed time series analysis procedure.

Chapter 7

Experimental Validation

Work from this chapter has been published in [183, 57, 204]

7.1 Introduction

The results of the literature survey presented in Ch. 2 have shown that there was a lack of R_D investigations comparing the results of data analysis to standardized testing. For this reason, an experimental approach was designed, in order to extract as much information as possible about the phenomenon of degradation. In this approach, the evaluation is based on indoor testing using international standards [141] and is by definition performed ex situ.

In order to estimate degradation through standardized testing, the capacity degradation rate, R_{D_C}, quantity was defined. R_{D_C} represents the degradation rate of the P_{nom} of a PV module/array measured at distinct points in time at STC.

> **Definition (Capacity degradation rate):** A scalar quantity which is defined as the annual, linear percentage reduction of the P_{nom}, measured at STC.

The experimental protocol was defined using the standard industry practice where the first set of module ratings at STC are measured as part of the PV plant pre-commissioning phase. The initial ratings are typically required by the PV plant owner/financer for assurance and by the installer for provision of warranties. Subsequent module/array ratings at STC can be performed at any point in time after commissioning, following some amount of field exposure. A significant amount of manual labour and system downtime is thus required, in order to dismount a random subset of modules from the array and test them in a solar simulator [139], either in an accredited standalone laboratory or in mobile laboratories which have become very popular recently because of the very risk of module damage from transportation to the standalone laboratory. Testing at STC is also usually combined with IR, as well as EL imaging to identify the manifestation of degradation mechanisms and the causes of under-performance. In order to evaluate the long-term R_{D_C}, the electrical characteristics prior to exposure or post-stabilization are compared with the measured characteristics after outdoor exposure and are used to calculate the percentage rate of change.

With respect to warranties, this rate of change of the P_{MPP} measured at STC can be converted to an annual value, which would represent the annual performance rate of change or R_{D_C}, in units of %/y.

7.2 Initial degradation

In order to study the initial degradation of different technologies, new PV modules from three different manufacturers and technologies were deployed in the field in both a system and standalone configurations. The three systems were designed with $1\,\mathrm{kW_p}$ capacity and identical BOS components, whereas the standalone modules were connected to outdoor continuous IV characterization systems side-by-side with their system-configured counterparts. The technologies of the modules under study were poly-Si, a-Si and CIGS with their characteristics listed in Table 7.1.

Table 7.1: Deployed PV systems and modules.

Manufacturer	Model	Technology	Installed on	Total Exposure [months]
Schott Solar	ASI103	a-Si	2012-07-09	47
T-Solar	TS95	a-Si-2J	2012-07-09	7
TSMC	TS150	CIGS	2014-01-30	30
QCells	Q.PRO-G3 255	poly-Si	2014-06-01	24

The standalone modules were periodically dismounted from the POA and measured at STC inside the solar simulator (every two to three weeks on average.) From the periodic indoor testing of the a-Si modules at STC and through analysis of the measurements, it was observed that the measured P_{STC} exhibited seasonal behavior, even when the measurement environment (STC) was stable [204]. Fig. 7.1 shows this seasonal behavior, after subtracting the mean, $E[P_{STC}]$, and normalizing with respect to the P_{nom}. On the same figure, the linear fit is presented, as a blue line.

Since the indoor testing conditions were stable across all chronological measurements at STC, this metastable behavior was attributed to module degradation [205], the Stæbler-Wronski effect (SWE) [19] and thermal annealing [206]. Degradation was assumed to be linear, since the first four months of outdoor exposure were omitted from the analysis and the outdoor performance had been stabilized. In this regard, the annual degradation rate was calculated at 2.36 %/y (±0.31) at a 95 % confidence level, where the uncertainty was calculated from the standard error of the slope estimator [184].

Finally, the results of the experimental investigation of the high initial degradation, correlate well with the results of the segmented R_{D_E}, estimated as described in Sec. 6.4.2.

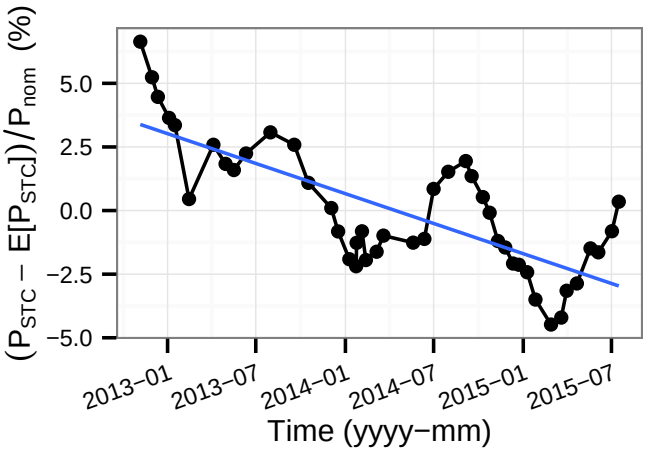


Figure 7.1: De-meaned and normalized module power measured indoors at STC, at regular intervals.

7.3 Reversible degradation

7.3.1 Light-induced metastability

Particularly for a-Si, the field performance features metastabilities which were described by the SWE in the literature [19].

In order to quantify this percentage of reversible degradation, the following procedure was devised: Firstly, the measured P_{STC} was normalized with respect to the P_{nom} and then the linear fit representing degradation was subtracted from the result. This produced the seasonal component of Fig. 7.2, which had a mean of zero and variance, $\sigma^2 = 2.71$.

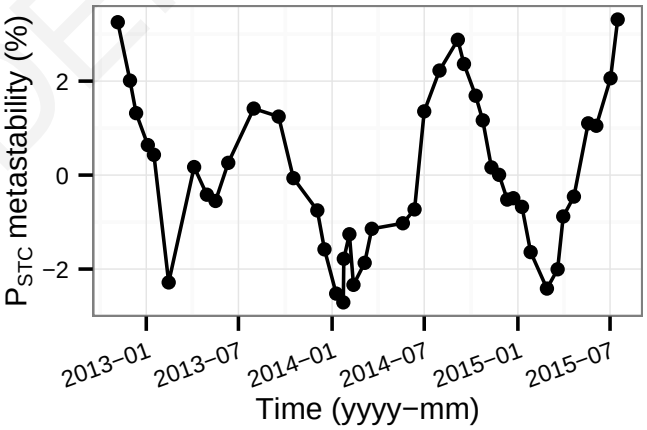


Figure 7.2: Residual metastability of P_{STC}, after normalizing to the P_{nom} and subtracting the linear degradation rate.

The effects of the module's metastable behavior on the field performance were analyzed in conjunction to indoor measurements at STC. It has been shown in Sec. 7.2 that the module had suffered a degradation rate of 2.36 %/y, while the SWE and thermal annealing resulted in ±3% variation of the performance, observable throughout the whole year.

Fig. 7.3 shows the normalized and temperature corrected P_{MPP} from the field and the normalized P_{STC}. It can be seen that there were differences between the two performance metrics, resulting in a mean absolute percentage error (MAPE) of 2.9 % for the whole evaluation period, calculated through interpolation of the discrete P_{STC} ratings and comparison to the filtered and corrected P_{MPP} from the field. It is also evident that the filtered and corrected P_{MPP} was higher than the normalized P_{STC} during summer and lower during winter.

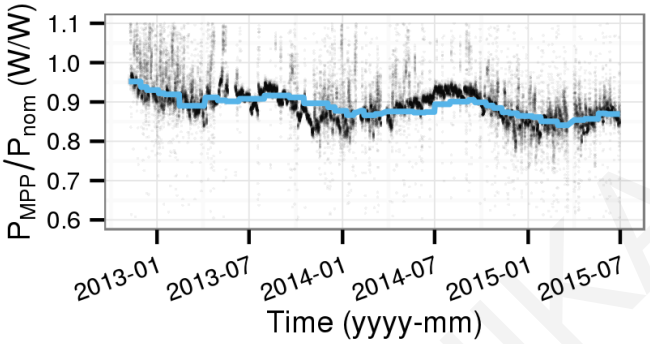


Figure 7.3: Normalized and temperature corrected P_{MPP} from the field (black colour) and normalized P_{STC} (blue line).

7.3.2 Soiling

Previous studies

Soiling can have a significant effect on PV systems whose modules are not regularly cleaned. Soiling and the quantification of its effect on the energy yield of PV generators presents a very challenging problem, due to its unpredictability. Some approaches have tried to deal with soiling by training artificial neural networks (ANNs) on measurements from clean PV modules and classifying large shifts in performance to soiling [207]. Other research has shown that soiling losses could be correlated with instances of rainfall and the size of dust particles in the vicinity of the PV generator [208], although specialized sensors were required on the field.

It has also been shown that soiling involved a complex interaction of dust and rain [209] and that there was a marked decrease in module efficiency during the dry season. Similarly, systems located in deserts were very susceptible to dust build-up, with performance loss ranging from 0.1 %/d to 0.3 %/d. Meanwhile, systems receiving frequent rainfall (at least once a month) did not experience significant losses. The average annual loss, according to this model, ranged from 1.5 % to 6.2 %, a figure which was most commonly reported in the literature [210].

In addition, it has been estimated that soiling could account for up to 2 % of the annual energy yield variability [211]. This effect was found to be dependent on the site, climate, human activities nearby and on flora and fauna.

Experimental investigation

Throughout this work, soiling has been considered to be a latent effect in PV performance. The very gradual performance loss due to soiling could be easily confused for degradation. For this reason, all PV modules and arrays under test had been on a regular cleaning schedule to minimize energy yield losses from soiling, as much as possible. This is also in-line with realistic PV deployment scenarios, where the installer and/or owner of the plant is responsible for its maintenance. It is expected that the performance of PV plants which are left unmaintained will eventually drop due to soiling, by an amount that is easily distinguished from typical generation profiles by a bump in the performance, immediately after cleaning.

Even though it was not the main concern of this study, since all modules in the test field were cleaned on a regular basis, an experimental evaluation of the effect was performed for the climate in Cyprus. The experiment consisted of measuring the difference in the P_{STC} of a-Si PV modules, before and after short-term outdoor exposure with no cleaning performed in between.

The specific a-Si modules used in this study were pre-conditioned and were past their initial degradation phase through prolonged exposure. This was verified from P_{STC} stability prior to performing this experiment. In addition, to minimize the effect of the SWE, the experiment was performed throughout the summer, when the constant high temperatures in Cyprus would have annealed the a-Si and mitigated the effect of SWE. This was again confirmed by stability of the P_{STC} measured indoors.

The results have shown that the P_{STC} of the modules had dropped by up to 10 % during the three-month outdoor exposure period. The procedure consisted of dismounting the PV module from the array and testing it at STC without cleaning it. The module was then cleaned thoroughly and retested at STC a second time. The percentage difference in the two sets of measurements represented reflection and recombination losses due to the layer of dirt on the front surface of the module.

In conclusion, this standalone experiment has showcased the importance of PV module cleaning. Experimental results in the climate of Cyprus cannot be extrapolated to other climate zones, or even other micro-climates inside Cyprus's climate zone, due to random factors in the vicinity of the PV testing site and weather unpredictability. For example, a PV plant installed alongside a popular high-way will require more frequent cleaning than one installed in a residential area. In both cases though, the effect will be large enough to influence any degradation rate estimates if left untreated for a long time.

7.4 Capacity degradation rate

7.4.1 Indoor testing at Standard Test Conditions

To validate the results of the analysis of field measurements presented in Ch. 3, Ch. 5 and Ch. 6, all PV arrays were completely disassembled and tested indoors at STC inside the solar simulator, in order to measure the electrical characteristics of each PV module using standardized test procedures and guidelines [35]. The electrical characteristics were then used to calculate the array nominal degradation rate. This indoor measurement procedure with two temporal STC ratings was well represented in the literature with a number of real world studies related to the estimation of the R_{D_C}.

All c-Si PV modules were characterized at STC, using the reference modules of the laboratory to ensure the calibration of the solar simulator, as described in Sec. 3.2.1. The expanded combined uncertainty of the indoor measurement at the 95 % confidence level was calculated at ±3.5 % for the P_{MPP}, as in Sec. 3.2.

An additional testing step was required for ensuring correct measurement of the ucy05 heterojunction with intrinsic thin-layer (HIT), ucy12 CIGS, ucy13 CdTe and ucy14 a-Si modules, as their spectral response was different from the spectral response of c-Si cells and subsequently, the calibration factor of the solar simulator had to be adjusted to account for the MMF. The spectral MMF of these technologies was calculated through outdoor calibration of the best performing module of each type at global AM1.5 in October 2014 in Cyprus [212]. The MMF was then used to correct the flasher's reference cell sensitivity, separately for each PV technology.

To calculate the MMF, the modules were connected to MPP tracking IV tracers for five days in order to acquire a clean set of data at high irradiance and clear sky. This procedure introduced additional uncertainty due to measurement, the data acquisition system and the irradiance and temperature sensors. The uncertainty components of the outdoor AM1.5 calibration procedure are listed in Table 7.2. Using this information, the uncertainty of the

Table 7.2: Uncertainty components of the AM1.5 calibration.

Device	Model	Uncertainty
Data logger	Papendorf	±0.01 %
Electronic load	ISET-mpp	< ± 1 %
Pyranometer	Hukseflux SR11	Dir. error: ±20 W/m^2, Non-linearity: ±1 %, Instability: < ± 1 %/y
Reference cell	ISET mono	< ± 4 %
Module temperature	Pt1000	DIN B accuracy class

MMF at 68 % confidence level, u_{MMF}, was calculated as ±3.3 % using typical uncertainty

propagation techniques [128] as follows:

$$u_{MMF} = \sqrt{u_{SR11}^2 + u_{mono}^2 + u_{DL}^2 + u_{load}^2 + u_{temp}^2} \tag{7.1}$$

where u_{SR11} is the uncertainty of the pyranometer, u_{mono} is the uncertainty of the mono-Si reference cell, u_{DL} is the uncertainty of the data logger, u_{Load} is the uncertainty of the electronic load and u_{temp} is the uncertainty of the temperature sensor installed on the back of the module.

The uncertainty of the initial maximum power prior to exposure, P_0, measured by the manufacturers at STC, $u_{P_{A_0}}$, was due to the power tolerance, u_{tol}, and the standard deviation of the flashing results from the manufacturer of each module, u_{T_0}:

$$u_{P_{A_0}} = \sqrt{u_{T_0}^2 + u_{tol}^2} \tag{7.2}$$

Similarly, the uncertainty of the array P_{STC} after 101 months of outdoor exposure, $P_{A_{101}}$ and subsequently $u_{P_{A_{101}}}$, was calculated by combining the flasher uncertainty, $u_{flasher}$, the standard deviation of repeated measurements for each PV module, u_{rep}, and the uncertainty of the MMF, where applicable, as in Eq. 7.3:

$$u_{P_{A_{101}}} = \sqrt{u_{flasher}^2 + u_{rep}^2 + u_{MMF}^2} \tag{7.3}$$

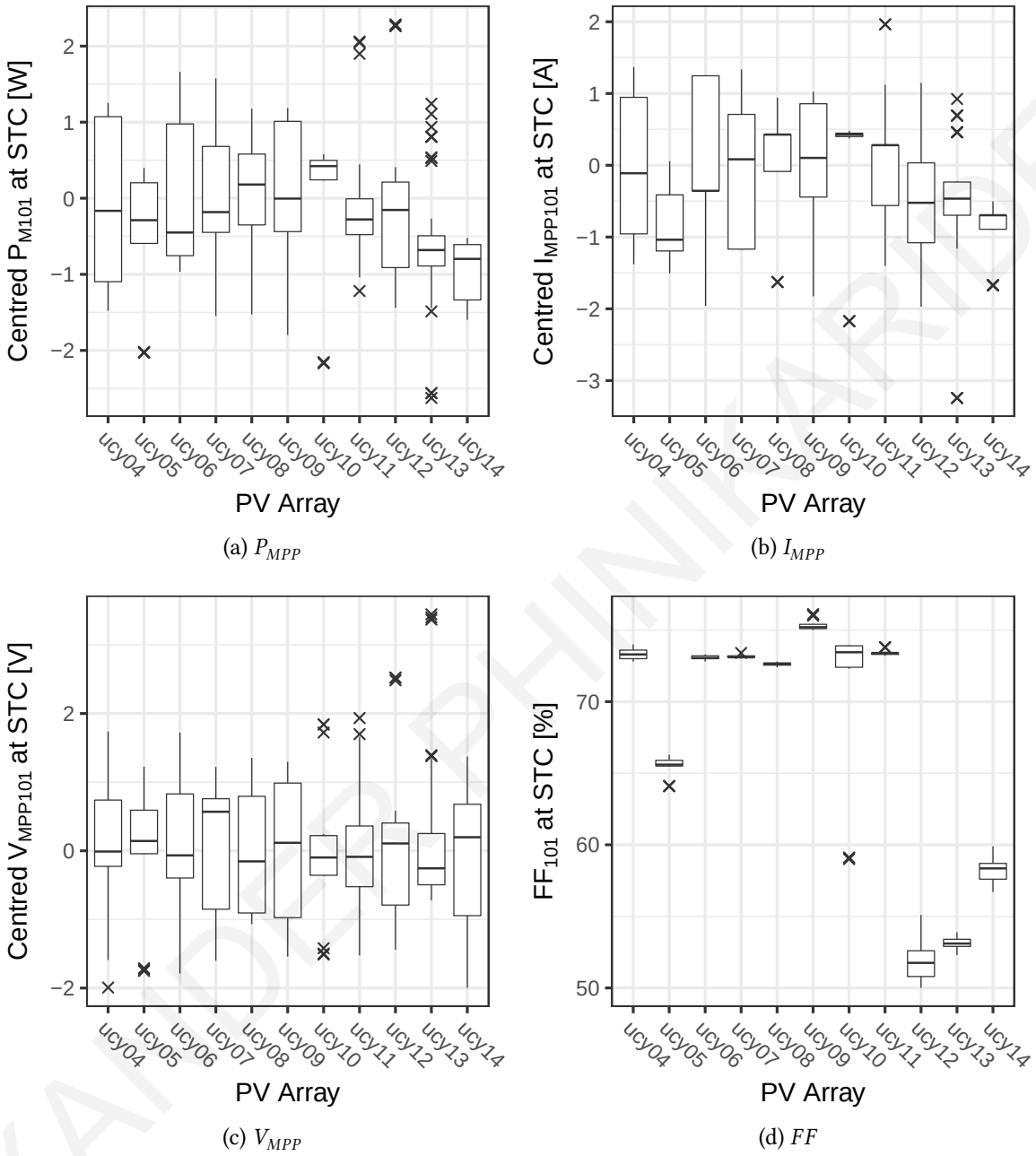
7.4.2 Module mismatches

The results of indoor testing at STC have shown significant variation in module P_{MPP} after 101 months of outdoor exposure, $P_{M_{101}}$, within the same array (more than 10 W in some cases) as shown in Fig. 7.4, which were mostly due to differences in the I_{MPP}. The variation on the V_{OC}, and the I_{SC}, of identical modules was well within the experimental uncertainty for c-Si, even for modules with visual defects.

Through analysis of variance (ANOVA) [213], it was shown that the variance of the $P_{M_{101}}$ of all modules tested at STC was strongly dependent on factors describing the model of each module and the serial number of each module. The "module model" factor resulted in a p-value near zero, with RSS of 3208, whereas the "serial number" factor also resulted in a p-value near zero, with a RSS of 1. The statistically significant ANOVA results suggested that the null hypothesis H_0 that the mean $P_{M_{101}}$ was the same across all modules in a PV array must be rejected.

These values have proven that there were two significant factors which contributed to the variability of the measured $P_{M_{101}}$, namely the "module model" and "serial number" and additionally, the low RSS of the "serial number" factor indicated that the variance of the measured $P_{M_{101}}$ could be fully explained by the characteristics of each individual PV module. Based on this, for the calculation of the array capacity degradation, each module was treated as an individual component and the P_{STC} for the whole array, $P_{A_{101}}$, was calculated by taking into account the array circuit topology and the position of each module in

the array. The same procedure was also used to calculate the initial array P_0, P_{A_0}, from manufacturer flash test reports.



(a) P_{MPP} (b) I_{MPP}

(c) V_{MPP} (d) FF

Figure 7.4: Variability of the centred PV array characteristics, measured through indoor testing at STC on all array modules after 101 months of field exposure.

As all PV modules in the arrays were connected in series, current mismatches were the most common type of mismatch encountered. The simple approach of accounting for the lowest current flowing through a string of modules was employed to construct the $P_{A_{101}}$. The total voltage produced by the string was taken as the sum of the individual module voltages as the differences in the V_{OC} due to the logarithmic dependence to I_{SC} were negligible. The consequence of mismatch loss was that the total power output of the PV arrays was lower than the sum of the individual power outputs of the modules.

7.4.3 Array capacity degradation rate

Using the initial MPP power of each PV array at STC, P_{A_0}, and the array MPP power at STC after 101 months of outdoor exposure, $P_{A_{101}}$, and modelling a linear slope through the two points in time, the annual array R_{D_C} was calculated as follows:

$$R_{D_C} = \frac{P_{A_{101}} - P_{A_0}}{P_{A_0}} \times \frac{100}{101/12} \tag{7.4}$$

Finally, the combined uncertainty of the degradation rate at STC, u_{RD}, was calculated by using Eq. 7.4, $u_{P_{A_0}}$ and $u_{P_{A_{101}}}$, with the following formula:

$$u_{RD} = \sqrt{\left(\frac{\partial R_D}{\partial P_{A_{101}}} \times u_{P_{A_{101}}}\right)^2 + \left(\frac{\partial R_D}{\partial P_{A_0}} \times u_{P_{A_0}}\right)^2} \tag{7.5}$$

The expanded uncertainty at a 95 % confidence interval was then calculated as in Eq. 6.12, with $k = 2$.

The results are shown in Fig. 7.5, side-by-side with the results of field data analysis described in Ch. 4, Ch. 5 and Ch. 6. From the figure, very good agreement can be observed for all PV systems under study, within the confidence intervals, except ucy09, ucy10 and ucy12. To investigate the differences within these systems, additional indoor characterization of all modules was performed.

7.4.4 Non-destructive characterization

Additional non-destructive characterization was performed in order to gain insight into the causes of physical degradation of the modules. All modules which were measured at STC were also imaged via EL, in order to identify cell defects not visible to the naked eye. IR thermography was also employed before dismounting the modules to check for hot spots [214].

Major defects in the form of broken cells and cracks were discovered for the ucy09 and ucy10 module technologies, as shown in Fig. 7.6 and Fig. 7.7. These EL images show modules with severely cracked and broken cells, one of which was caused by a hot spot. The modules were operating at two thirds capacity when tested inside the solar simulator, which can be seen by the large outliers and variability of the *FF* in Fig. 7.4. The EL images had thus provided proof towards the low performance of some of the ucy09 and ucy10 modules measured indoors at STC.

Other arrays with low performing modules, such as ucy05, ucy12 and ucy13 were also found to have a high R_{D_C}. Common between these modules was the fact that they were all made with thin-film layers. In the case of ucy05, the a-Si thin-film layer was intrinsic, between the *p* and *n* Si layers, and this was captured by EL. Fig. 7.8 shows the EL image of the best performing module from the array. Even though they are small in size, the dark areas correspond to shunts and do not produce any current. For comparison, EL images of

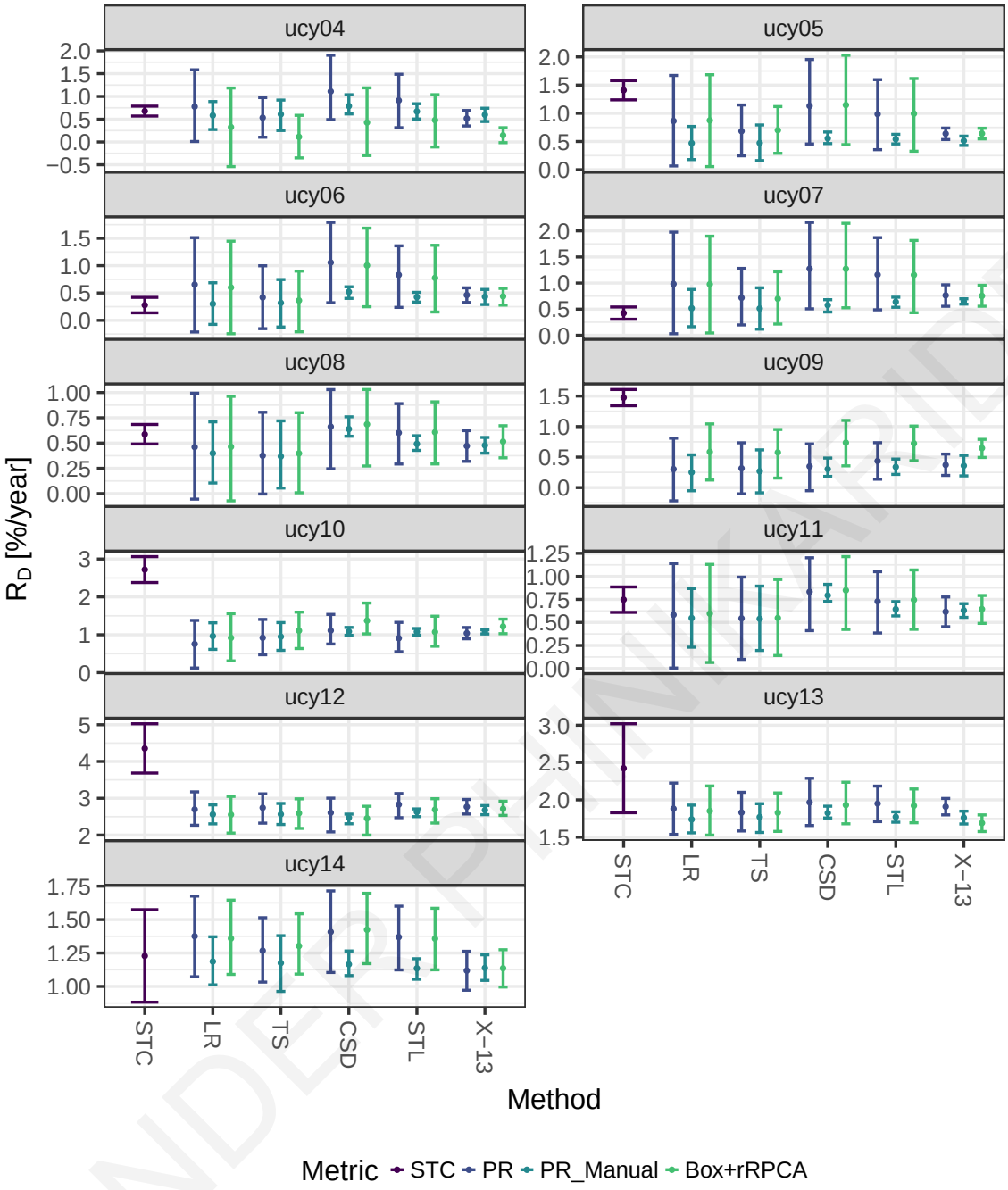


Figure 7.5: Annual energy and capacity R_D evaluated through analysis of field performance metrics and indoor testing at STC.

problematic modules from the ucy05 array are shown in Fig. 7.9. The images reveal deeper physical defects and whole cells lost. In addition, some kind of checkerboard pattern common to all modules can be identified. The existence of a pattern can be attributed to manufacturing, since all modules were from the same batch. Inspecting the images closer can reveal that some shaded cells were not completely dead; in fact, the most probable explanation of why some of them appear black could be the intrinsic a-Si layers underperforming and limiting the current through the cell.

In the case of purely thin-film arrays, such as ucy12, ucy13 and ucy14, the high estimated R_{D_C} were in line with what was found in the literature (see Ch. 2.) An EL signal

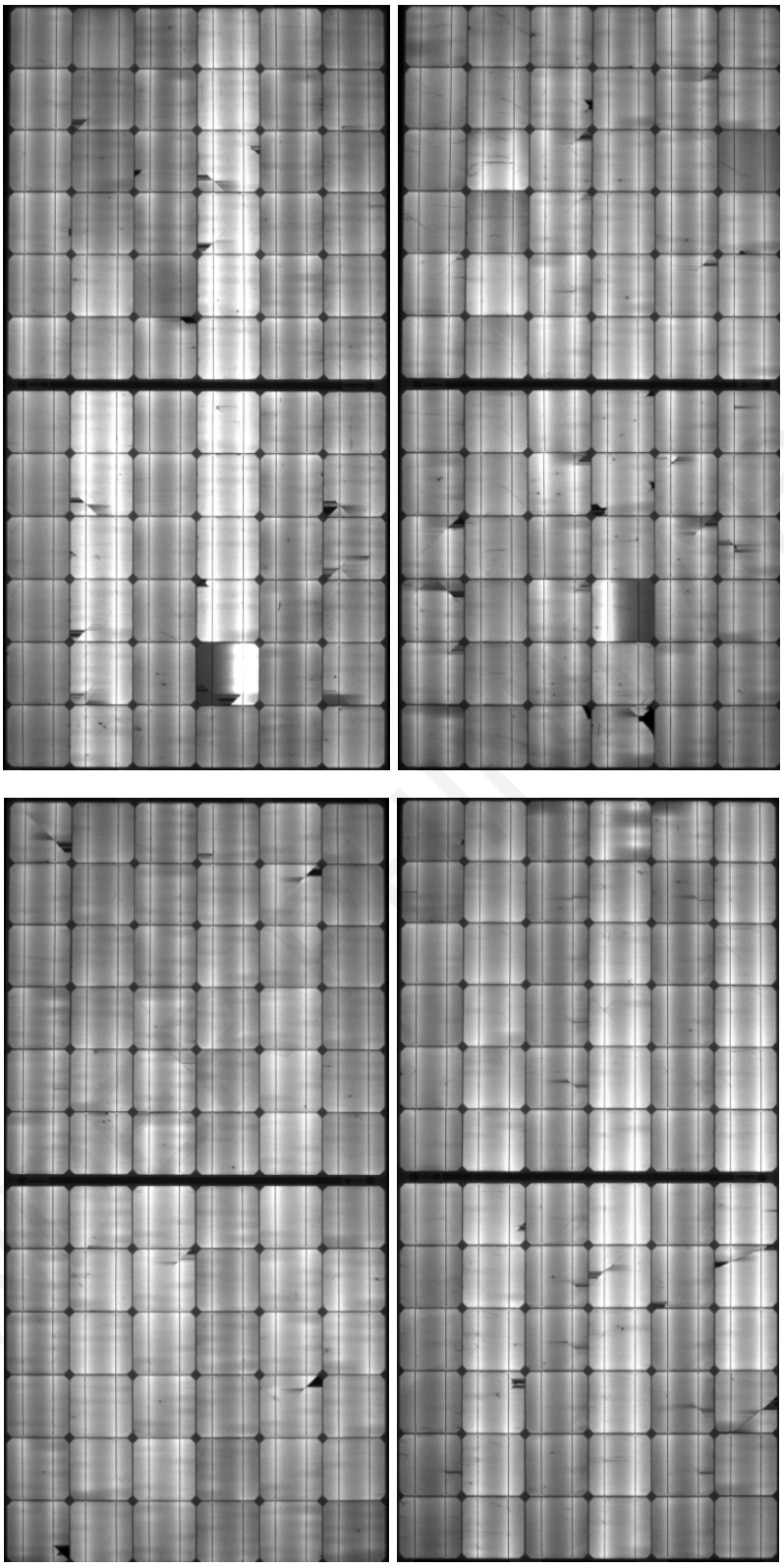


Figure 7.6: EL images of four problematic modules from the ucy09 PV array.

was more difficult to detect for these technologies, apart from ucy12 which was CIGS technology and its EL wavelength was close to c-Si, therefore it could be captured effectively by the EL camera used. The ucy13 modules required high bias voltage to emit any kind of EL signal, therefore EL was performed on only a few modules to avoid damaging them.

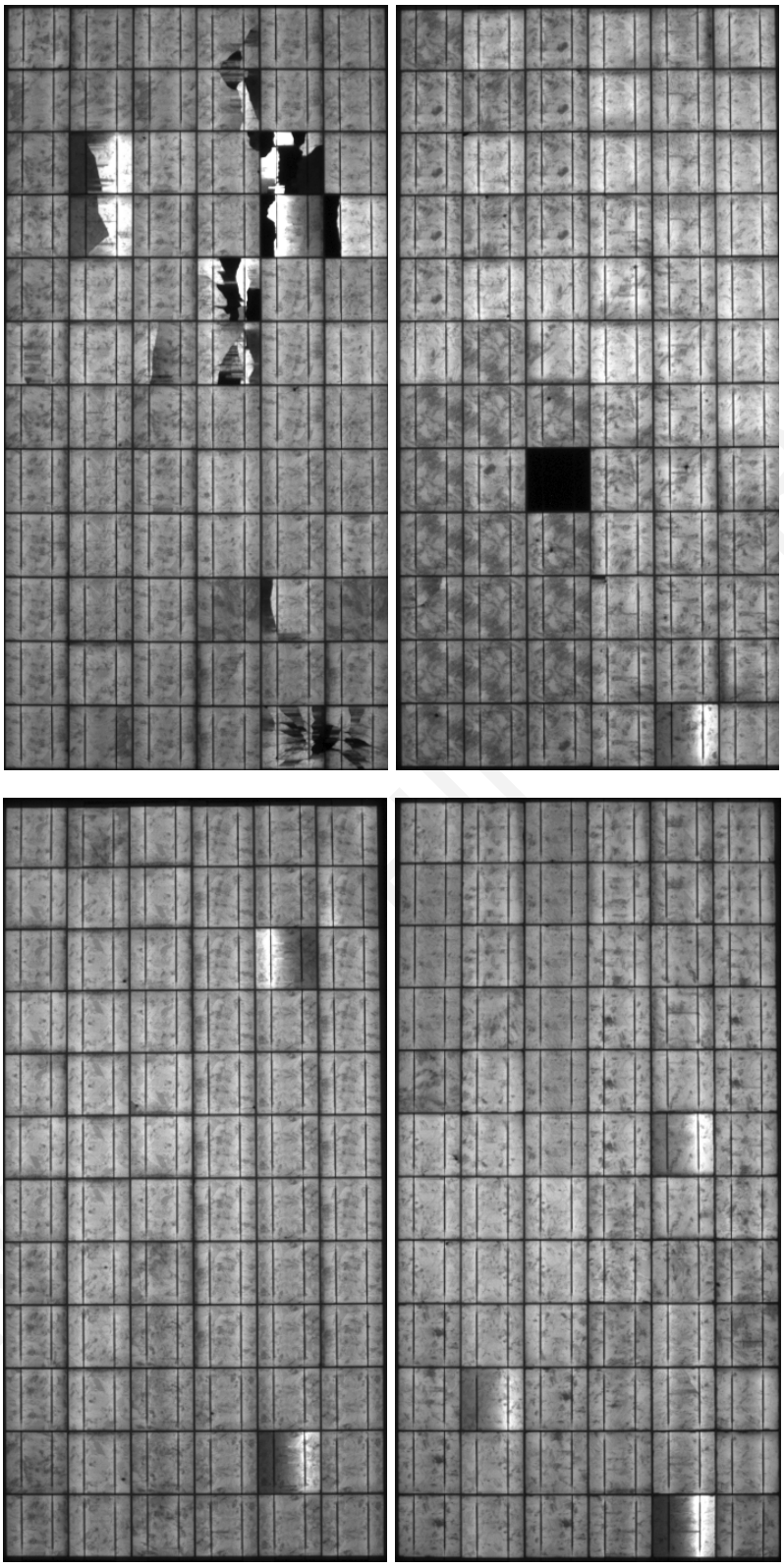


Figure 7.7: EL images of four problematic modules from the ucy10 PV array.

Regarding ucy14 modules, a bias voltage could not be applied properly and therefore, no EL images could be recorded.

Fig. 7.10 and Fig. 7.11 show typical EL images of these modules. On ucy12 modules, black vertical lines can be distinguished which correspond to shunted cells (the cells are

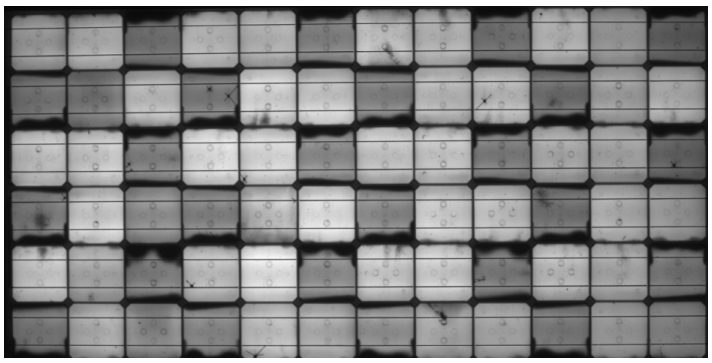


Figure 7.8: EL image of the best performing module of the ucy05 PV array.

long and thin.) In addition, degradation of the transparent conductive oxide (TCO) can be seen by the halo-like appearance and patterns that appear as swishes. Also, there was pronounced degradation around the edges of the modules.

Finally, an EL image of a typical module from the ucy13 array is shown in Fig. 7.11. As with ucy12, but less pronounced, halos and degradation around the edges can be distinguished.

7.5 Comparison of analysis methods

The final results of R_{D_E} and R_{D_C}, shown together in Fig. 7.5 for all PV arrays, analysis methods and performance metrics, demonstrate that for PV arrays with no physical defects the estimated R_D was comparable between analysis of filed measurement data and ex-situ characterization at STC, therefore validating the approach.

On the performance metrics dimension, the uncertainties of the R_{D_E} estimated from uncorrected monthly PR and corrected monthly PR^*_{rRPCA} were comparable across all systems and all analysis methods, with the uncertainty from PR^*_{rRPCA} begin equal or slightly less than from uncorrected PR. The lowest uncertainty from the metrics on the R_{D_E} was achieved by using the manually corrected PR which was tedious to produce. For all c-Si arrays, the uncertainty of the R_{D_C} using STC measurements was the lowest, as expected. Such low uncertainties on the R_{D_E} were only achieved by using X-13ARIMA-SEATS with any performance metric. On the contrary, due to spectral mismatch, the corresponding uncertainty of all thin-film arrays (ucy12, ucy13 and ucy14) was higher in the indoor laboratory than the field.

On the analysis method dimension, it can be seen that due to optimal seasonal decomposition, the X-13ARIMA-SEATS method resulted in the least amount of uncertainty on R_{D_E} across all systems and performance metrics. The second lowest uncertainty on the R_{D_E} was achieved by using the TS method. The third lowest uncertainty was achieved by using STL, the fourth by using CSD and finally, the highest uncertainty was exhibited by LR which was most widely used in the literature.

Regarding the magnitude of the R_D, it can be seen in Fig 7.5 that R_{D_C} results overlapped R_{D_E} results only for PV arrays with no problematic modules. The highest differences were

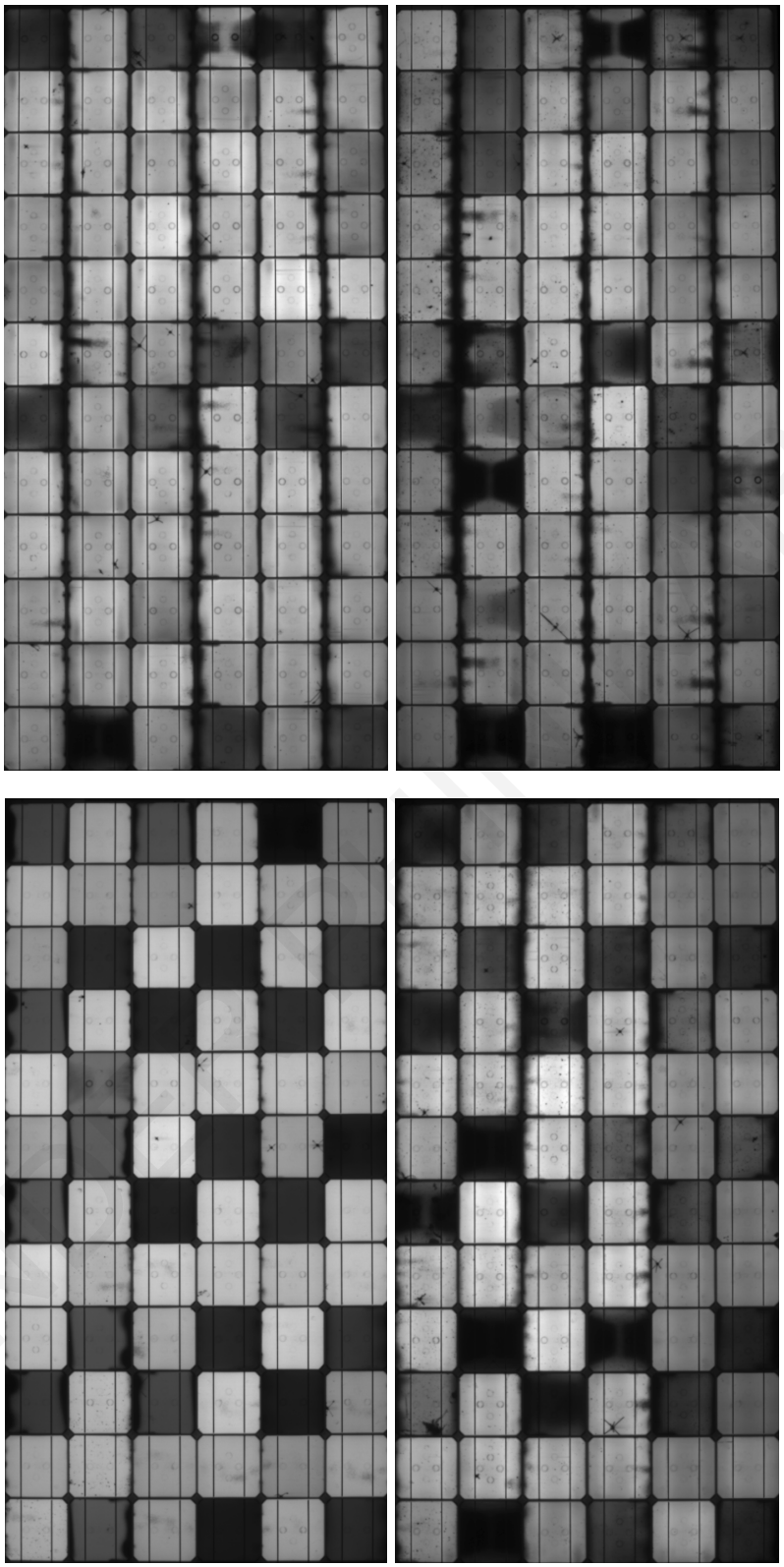


Figure 7.9: EL images of typical problematic modules from the ucy05 PV array.

recorded for the ucy09, ucy10 and ucy12 systems which had suffered the worst physical damage and degradation. Between analysis methods, the mean as well as the uncertainty varied with different performance metrics. The R_{D_E} from uncorrected *PR* presented the highest deviation from R_{D_C}. Treating the data with outlier filters and missing data impu-

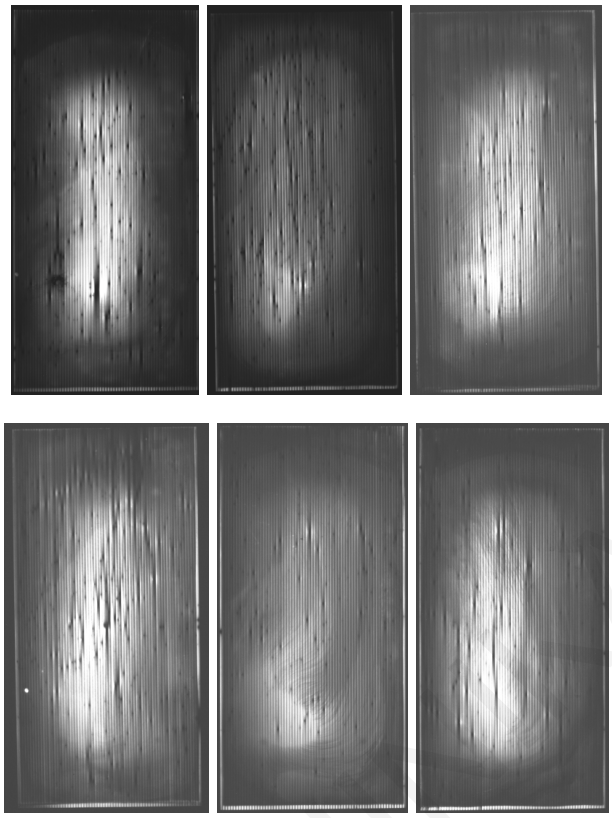


Figure 7.10: EL images of typical modules from the ucy12 PV array.

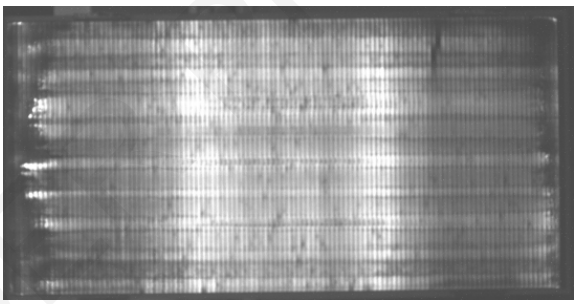


Figure 7.11: EL image of a typical module from the ucy13 PV array.

tation to produce PR^{*}_{rRPCA} lowered the deviation. As expected, the manually corrected PR was the most agreeable to STC, for arrays with healthy modules.

Specifically for PR^{*}_{rRPCA} metrics. especially strong agreement has been demonstrated between STC, X-13ARIMA-SEATS, TS for healthy arrays. This is evident from the overlap of the R_D values. A major advantage of X-13ARIMA-SEATS and TS was that the uncertainty of the methodology was much lower in comparison to other methods, providing increased confidence in the estimated annual R_D. Weaker agreement was exhibited by the CSD methodology.

To better quantify differences between R_{D_E} and R_{D_C}, the mean absolute percentage deviation (MAPD) was computed for each combination of PV system and trend estimation

method. The MAPD was given by Eq. 7.6:

$$MAPD = \frac{100}{n} \sum_{i=1}^{n} \left| \frac{R_{D_C} - R_{D_E}}{R_{D_C}} \right| \tag{7.6}$$

where n was the number of performance metrics, i.e. PR, PR with manual corrections and PR^*_{rRPCA}. The MAPD is plotted in Fig. 7.12. The MAPD quantifies and highlights more

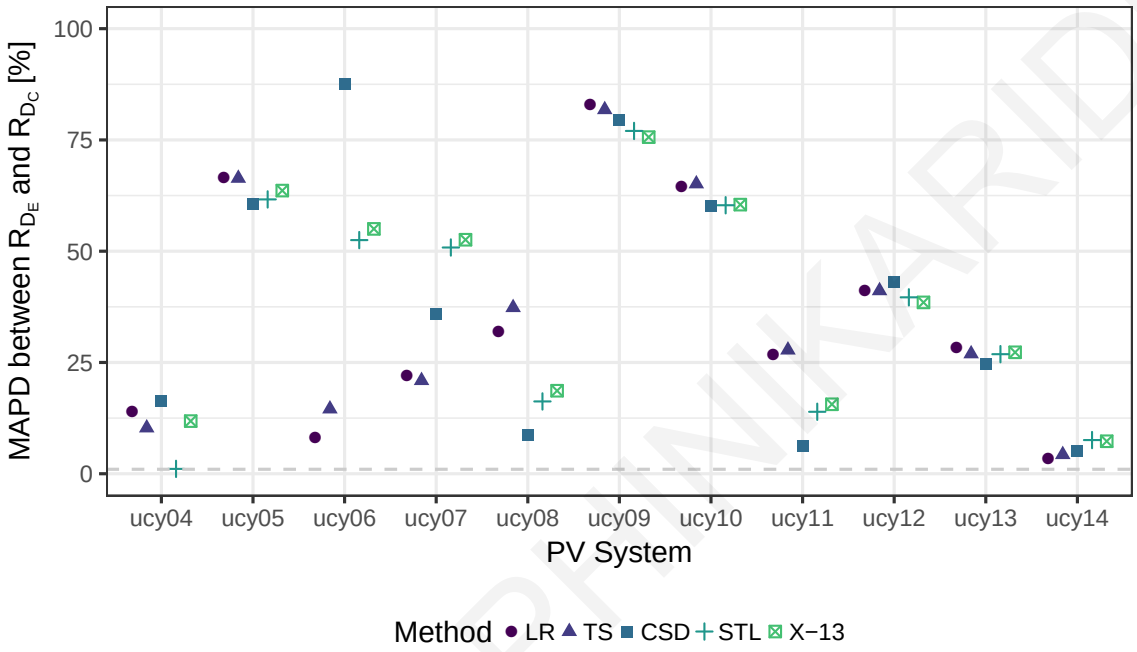


Figure 7.12: Mean Absolute Percentage Deviation (MAPD) between R_{D_E} and R_{D_C}.

clearly differences between analysis of field measurement data and testing ex-situ. Highest differences were observed for the ucy09, ucy10 and ucy05 array, which can be traced back to physical defects (see Sec. 7.4.4.) In addition, an outlier can be seen for ucy06, using CSD to extract the trend. Finally, except ucy06 and ucy14, the differences between R_{D_C} and R_{D_E} were reduced by using the time series analysis methodologies developed in this work instead of LR.

Chapter 8

Conclusions

8.1 Degradation rate estimation in photovoltaics

Through this work, it has been shown that the subject of degradation in PV poses many interesting challenges. The methodologies developed in this work, through the application of generalized statistical analysis methods on time series of measurements from PV systems in the field have shown that an unsupervised and generalized methodology could be developed for assessing the degradation rate, without having to disrupt normal system operation.

The approaches of field data analysis and indoor testing at STC, as presented in Ch. 3 and Ch. 7 were applied on eleven different c-Si and thin-film grid-connected PV plants, operating side-by-side at the PV Technology test site of the University of Cyprus since June 2006. More specifically, P_A from grid-connected PV arrays and G_I, from calibrated pyranometers on the POA were used to derive an accurate estimation of the energy degradation rate, R_{D_E}, and provide statistical inference on the results. The results were supported by the more comprehensive and hands-on approach of indoor testing at STC according to international standards. The indoor procedure provided extensive detail of the characteristics of each module in the array, the major mechanisms behind the observed performance degradation and additionally, it carried the least amount of uncertainty but required a significant amount of manual labour, introduced risk of module damage due to handling and required significant system downtime.

The developed data analysis methodology was designed to provide data qualification, outlier detection and rejection, performance metric creation, seasonal decomposition and linear/non-linear trend modelling. The most interesting feature of the data analysis approach was that no prior knowledge of the physical properties or the circuit of the PV system was required, except for the datasheet power, making this approach a very good candidate for an unsupervised algorithm and integration into operations and maintenance (O&M) systems.

In addition, the comparison between the results of outdoor data analysis and indoor testing at STC has shown good agreement between the two methods in the case of non-

problematic modules in the array. In the case of problematic modules in the array, there were differences in the estimated R_D. Through extensive indoor characterization, the differences were attributed to physical damage on individual modules, such as cracked or broken cells, material degradation, TCO corrosion, browning of the EVA, hot spots and delamination, and to the spectral mismatch of non-c-Si module technologies. The differences in means between indoor module parameters in problematic arrays were large and affected the ability of the whole array to perform well. For arrays with only healthy modules, the mean R_{D_E} was within the confidence interval of R_{D_C}.

Between module technologies, the results have shown that arrays with modules containing a thin-film layer degraded at a higher rate than purely c-Si modules. Especially in this work which used modules which had been exposed for nine years, this high degradation was able to be observed both through the measurement data and through physical defects visible with EL imaging. Specifically, the R_{D_E} of arrays with thin-film modules ranged between 1.3 %/y to 4.5 %/y which was commonly reported in the literature. Physical degradation was easily observable through EL which revealed severe cases of shunted cells, TCO corrosion and module edge degradation. On the other hand, the R_{D_E} of arrays with healthy c-Si modules ranged between 0.25 %/y to 0.8 %/y, depending on the analysis method and performance metric. This was consistent with the literature as well. Whereas for the c-Si arrays the main factor was the analysis method and performance metric, for thin-film technologies the main factor was the actual cell and module technology. It was thus concluded that different PV technologies could be sufficiently characterized by varying ranges of R_{D_E} and R_{D_C}.

Based on the strong agreement with the R_{D_C} estimated at STC, the low amount of uncertainty, the robustness to missing data and the statistically significant model produced, the X-13ARIMA-SEATS regARIMA time series analysis method and the non-parametric TS method were proven to provide the most benefit from all methods tested, for estimating the linear R_{D_E} of a PV array with no serious physical faults. In general, application of the unsupervised methodology on field measurement data yielded more predictable results (in line with what is reported in the literature) due to the lower measurement correction and extrapolation requirements (i.e. spectral mismatch, temperature uniformity), in contrast to the indoor testing procedure.

Lastly, in the case of PV plants with significantly damaged modules, the indoor testing approach has provided more insight into the physical causes of under-performance. In the case of module technologies with different spectral response than c-Si, such as thin-film and multi-junction technologies, the results of the comparison have demonstrated the need for additional capabilities of the indoor testing laboratory, such as spectrally matched reference cells and an InGaAs EL camera to capture the EL signal of CdTe, a-Si and HIT module more effectively. These additional capabilities will enable drawing concrete conclusions regarding the PV array degradation rate.

8.2 Research achievements

The main outcome of this work is an index of R_{D_E} that is robust to outliers and that can be estimated in an unsupervised fashion. A secondary aspect of this work is the online estimation of the R_{D_E}. The developed methodology is computationally tractable and can be applied on-line. This involved the translation of the developed methodologies to proof-of-concept form that was capable of performing "real-time" estimation of the R_{D_E} in software. The software implementation of the data pipeline was optimized for speed and abstraction of the underlying details of the PV entities (modules, arrays, systems etc.) into their own environment, using the split-apply-combine paradigm for data analysis.

Through this work, the following research objectives were achieved:

1. Development of an improved methodology for classifying and filtering outliers in field measurement data
2. Development of an improved methodology for handling missing data points and investigation of its sensitivity
3. Development of an improved methodology for extracting the trend from field measurement data
4. Development of an improved methodology for estimating non-linear trends in PV performance
5. Extensive indoor characterization of all PV modules from the plants under test and qualification of all observed degradation mechanisms
6. Comparison of the results of the trend extraction procedure to the results of indoor characterization and quantification of observed differences
7. Definition of calibration campaigns for the collection of accurate field measurement data from multiple PV plants of different technologies at the UCY and tracking of the sensitivity of each sensor over time
8. Quantification of the reversible degradation of a-Si technologies

By leveraging robust statistical techniques, the degradation rate of deployed PV plants was estimated with higher accuracy and lower uncertainty than current industry practices and additionally, it provided inference on the results. The methodology was designed to be unsupervised, which decouples human influence from the analysis and removes bias introduced by current practices through the arbitrary selection of filtering, regression and extrapolation procedures.

8.3 Innovation

Based on previous findings, this work has gone beyond the state-of-the-art and addressed the question whether P_A and G_I alone could be sufficient for a reliable estimation of PV degradation in the field.

To the best of the author's knowledge, no other research work has tried to combine all the developed data analysis steps to arrive to degradation rate estimates from raw measurement data that provide inference, are robust to outliers and were estimated through a generalized unsupervised algorithm. A very important fact is that significant effort has been put into creating a baseline of indoor ratings at STC to compare the feasibility and accuracy of the algorithm, which no other study has performed at such scale. Each one of the PV modules under study were dismounted from the arrays and extensively characterized indoors using a calibrated solar simulator and state-of-the-art visual techniques for identifying defects. Another important innovative result of this effort was that the metastable behaviour of a-Si PV modules was quantified and its reversible degradation was demonstrated and quantified as a result of the outdoor exposure.

Finally, since the methodology was benchmarked on PV plants of various technologies operating side-by-side for the long term, the methodology has the potential to be generalized across multiple PV technologies, module manufacturers and installation sites.

8.4 Future work

This dissertation has presented original work and results beyond the state-of-the-art in a subject that has gained a lot of interest in the past few years and is currently under much research in the field of PV. Taking advantage of the momentum created and the increasing need for data science in this field, the opportunity for improving upon the presented methodologies in the future is evident.

The first step for improving this work and developing it into a universally accepted methodology for estimating the degradation rate would be to benchmark the whole data pipeline with measurements from PV plants deployed outside of Cyprus. It is expected that this will present new challenges, as more types of PV panels and BOS will be encountered, to reveal new, unobserved up to now degradation modes and operational/seasonal characteristics. Although the developed methodology is generalized due to its reliance on statistical tests, the need for better statistical tests and procedures will surely arise, given the diverse conditions PV plants are commissioned in. For example, the estimation of degradation for PV arrays deployed in northern countries will present different challenges than the same PV arrays deployed in the desert.

Secondly, the use of different sources of irradiance measurements could be investigated. In case measurements from a PV plant were not recorded alongside an accurate irradiance sensor, the effects of using another source of irradiance measurements as a surrogate could be considered. The surrogate could either be a nearby irradiance sensor, a satellite provider

or another nearby monitored PV plant. In addition, the possibility of eliminating the irradiance sensor completely could be investigated, as it could result in simplifying the PV installation and reduce the uncertainty imposed by unmaintained and uncalibrated sensor equipment.

Thirdly, more advanced models could be used for modelling the PV performance time series. Growth curve models are another important type of latent variable models. Growth modeling could be used to analyze time series data, where a quantity is measured on several occasions, in order to study the change over time. This can then be modelled as a linear or non-linear curve. Linear dynamic models could also be investigated. These models can capture the temporal structure of a stochastic process and have been used in financial time series analysis. They can be used to analyze non-linear phenomena in a robust manner and can construct the long-term behaviour of the process from its time series. Both growth curve and linear dynamical models have similar characteristics to the observed degradation of PV modules.

Additionally, as demonstrated in Sec. 4.3, the probability distribution of PV system and meteorological measurements feature two distinct sub-populations. In this work, this effect was not dealt with due to the application of non-parametric and robust methods on the data. In future work, random and fixed effects models could be used to capture the information contained in each sub-population. An expected outcome of this investigation would be to model the two distinct modes of operation and to mitigate the effects of non-linear temperature and irradiance relationships with the P_A. This would improve forecasting and extracting the structure of the data under various environmental conditions.

Another topic that could be explored is the effect of cold-start, where not enough data are available in the early post-commissioning phase of the PV system to estimate the degradation. This poses challenges in the reliability of the estimation of degradation and the statistical power of the results. In the absence of enough measurement data for inference, Bayesian techniques could be explored, even though they are computationally much more expensive.

Finally, the topic of outlier and fault detection could be further developed. An immediate contribution could be made by formulating a better methodology for the results of PCA and RPCA and its variants, to distinguishing two separate subspaces, one of which will describe the normal operation and the other which will describe operation under fault under some measure of statistical confidence.

8.5 Articles in preparation

A number of research articles that stem from this work are currently in preparation:

1. Anomaly detection in PV with robust principal component analysis
2. Statistical estimation of shading losses for photovoltaic arrays

3. Effect of missing data on the analysis of photovoltaic system measurements
4. Generalized energy degradation estimation methodology for photovoltaics

Bibliography

[1] IRENA, "REthinking Energy 2017: Accelerating the global energy transformation," International Renewable Energy Agency, Abu Dhabi, 2017 (cit. on p. 2).

[2] D. C. Jordan, M. G. Deceglie, and S. R. Kurtz, "PV degradation methodology comparison — A basis for a standard," in *43rd IEEE PVSC*, IEEE, Jun. 2016, pp. 0273–0278, ISBN: 978-1-5090-2724-8. DOI: 10.1109/PVSC.2016.7749593 (cit. on p. 3).

[3] A. Phinikarides, N. Kindyni, G. Makrides, and G. E. Georghiou, "Review of photovoltaic degradation rate methodologies," *Renewable and Sustainable Energy Reviews*, vol. 40, pp. 143–152, Dec. 2014. DOI: 10.1016/j.rser.2014.07.155 (cit. on p. 6).

[4] M. Hadjipanayi, I. Koumparou, N. Philippou, V. Paraskeva, A. Phinikarides, G. Makrides, V. Efthymiou, and G. E. Georghiou, "Prospects of photovoltaics in southern European, Mediterranean and Middle East regions," *Renewable Energy*, vol. 92, pp. 58–74, Jul. 2016. DOI: 10.1016/j.renene.2016.01.096 (cit. on p. 6).

[5] J. H. Wohlgemuth, D. W. Cunningham, A. M. Nguyen, and J. Miller, "Long term reliability of PV modules," in *IEEE 4th World Conference on Photovoltaic Energy Conversion*, vol. 4968, 2006, pp. 57–58 (cit. on p. 6).

[6] J. H. Wohlgemuth and S. Kurtz, "Using accelerated testing to predict module reliability," in *37th IEEE PVSC*, 2011, pp. 3601–3605, ISBN: 978-1-4244-9965-6. DOI: 10.1109/PVSC.2011.6185927 (cit. on p. 6).

[7] U.S. Department of Energy, Sandia, and NREL, "Accelerated Aging Testing and Reliability in Photovoltaics," in *Solar Energy Technologies Program Workshop II*, 2008 (cit. on p. 6).

[8] N. Bosco, "Reliability concerns associated with PV technologies," National Renewable Energy Laboratory (NREL), 2010 (cit. on p. 6).

[9] D. C. Jordan, T. J. Silverman, B. Sekulic, and S. R. Kurtz, "PV degradation curves: Non-linearities and failure modes," *Progress in Photovoltaics: Research and Applications*, Jan. 1, 2016. DOI: 10.1002/pip.2835 (cit. on pp. 6, 15).

[10] D. C. Jordan, T. J. Silverman, J. H. Wohlgemuth, S. R. Kurtz, and K. T. VanSant, "Photovoltaic failure and degradation modes," *Progress in Photovoltaics: Research and Applications*, Jan. 1, 2017. DOI: 10.1002/pip.2866 (cit. on p. 6).

[11] A. M. Reis, N. T. Coleman, M. W. Marshall, P. A. Lehman, and C. E. Chamberlin, "Comparison of PV module performance before and after 11-years of field exposure," in *29th IEEE PVSC*, 2002, pp. 1432–1435, ISBN: 0-7803-7471-1. DOI: 10.1109/PVSC.2002.1190878 (cit. on pp. 6, 9, 17).

[12] P. Sánchez-Friera, M. Piliougine Rocha, J. Peláez, J. Carretero, and M. Sidrach de Cardona, "Analysis of degradation mechanisms of crystalline silicon PV modules after 12 years of operation in Southern Europe," *Progress in Photovoltaics: Research and Applications*, vol. 19, no. 6, pp. 658–666, Sep. 2011. DOI: 10.1002/pip.1083 (cit. on pp. 6, 7).

[13] S. Sakamoto and T. Oshiro, "Field test results on the stability of crystalline silicon photovoltaic modules manufactured in the 1990s," in *3rd World Conference on Photovoltaic Energy Conversion*, 2003, pp. 1888–1891, ISBN: 4-9901816-0-3 (cit. on p. 6).

[14] C. E. Chamberlin, M. A. Rocheleau, M. W. Marshall, and P. A. Lehman, "Comparison of PV Module Performance Before and After 11 and 20 Years of Field Exposure," in *37th IEEE PVSC*, IEEE, 2011, ISBN: 0-7803-7471-1. DOI: 10.1109/PVSC.2002.1190878 (cit. on p. 6).

[15] M. Quintana, D. L. King, T. J. McMahon, and C. R. Osterwald, "Commonly observed degradation in field-aged photovoltaic modules," in *29th IEEE PVSC*, 2002, pp. 1436–1439, ISBN: 0-7803-7471-1. DOI: 10.1109/PVSC.2002.1190879 (cit. on p. 6).

[16] Y. Hishikawa, K. Morita, S. Sakamoto, and T. Oshiro, "Field test results on the stability of 2400 photovoltaic modules manufactured in 1990s," in *29th IEEE PVSC*, 2002, pp. 1687–1690, ISBN: 0-7803-7471-1. DOI: 10.1109/PVSC.2002.1190944 (cit. on p. 6).

[17] C. R. Osterwald, A. Anderberg, S. Rummel, and L. Ottoson, "Degradation analysis of weathered crystalline-silicon PV modules," in *29th IEEE PVSC*, 2002, pp. 1392–1395, ISBN: 0-7803-7471-1. DOI: 10.1109/PVSC.2002.1190869 (cit. on p. 7).

[18] C. Radue and E. E. van Dyk, "A comparison of degradation in three amorphous silicon PV module technologies," *Solar Energy Materials and Solar Cells*, vol. 94, no. 3, pp. 617–622, Mar. 2010. DOI: 10.1016/j.solmat.2009.12.009 (cit. on p. 7).

[19] D. L. Staebler and C. R. Wronski, "Reversible conductivity changes in discharge-produced amorphous Si," *Applied Physics Letters*, vol. 31, no. 4, p. 292, 1977. DOI: 10.1063/1.89674 (cit. on pp. 7, 88, 89).

[20] M. Tayyib, T. O. Saetre, O.-M. Midtgaard, and J. O. Odden, "Analysis of Some Effects of Possible Device Degradation over a 30 Year Lifespan," in *26th EU-PVSEC*, 2011, pp. 1437–1441. DOI: 10.4229/26thEUPVSEC2011-2BV.2.13 (cit. on p. 7).

[21] C. R. Osterwald, T. J. McMahon, and J. A. del Cueto, "Electrochemical corrosion of SnO2:F transparent conducting layers in thin-film photovoltaic modules," *Solar Energy Materials and Solar Cells*, vol. 79, no. 1, pp. 21–33, Aug. 2003. DOI: 10.1016/S0927-0248(02)00363-X (cit. on p. 7).

[22] J. Siemer and C. Haase, "Stress relief?" *PHOTON Int. Magazine*, Jan. 2011 (cit. on p. 7).

[23] M. Köntges, S. Kurtz, C. Packard, U. Jahn, K. A. Berger, and K. Kato, "Review of failures of photovoltaic modules," International Energy Agency, IEA-PVPS T13-01:2014, 2014 (cit. on p. 7).

[24] A. Cronin, S. Pulver, D. Cormode, D. C. Jordan, S. R. Kurtz, and R. Smith, "Measuring degradation rates of PV systems without irradiance data," *Progress in Photovoltaics: Research and Applications*, Feb. 2013. DOI: 10.1002/pip.2310 (cit. on pp. 7, 17).

[25] D. C. Jordan, S. R. Kurtz, K. VanSant, and J. Newmiller, "Compendium of photovoltaic degradation rates," *Progress in Photovoltaics: Research and Applications*, 2016. DOI: 10.1002/pip.2744 (cit. on p. 7).

[26] M. Gostein and L. Dunn, "Light Soaking Effects on Photovoltaic Modules: Overview And Literature Review," in *37th IEEE PVSC*, 2011 (cit. on p. 7).

[27] C. R. Osterwald, J. Adelstein, J. A. del Cueto, B. Kroposki, D. Trudell, and T. Moriarty, "Comparison of Degradation Rates of Individual Modules Held at Maximum Power," in *IEEE 4th World Conference on Photovoltaic Energy Conversion*, May 2006, pp. 2085–2088, ISBN: 1-4244-0016-3. DOI: 10.1109/WCPEC.2006.279914 (cit. on pp. 7, 17).

[28] T. Nordmann, L. Clavadetscher, W. van Sark, and M. Green, "Analysis of Long-Term Performance of PV Systems," International Energy Agency, IEA-PVPS T13-05:2014, 2014 (cit. on p. 7).

[29] DNV GL, "Photovoltaic Module Degradation," DNV GL, RANA-WP-03-A, 2015 (cit. on p. 7).

[30] M. Woodhouse, R. Jones-Albertus, D. Feldman, R. Fu, K. Horowitz, D. Chung, D. Jordan, and S. Kurtz, "On the Path to SunShot: The Role of Advancements in Solar Photovoltaic Efficiency, Reliability, and Costs," National Renewable Energy Laboratory, Golden, CO, NREL/TP-6A20-65872, 2016 (cit. on p. 7).

[31] SunPower, "SunPower module degradation rate," 2015 (cit. on pp. 7, 13).

[32] A. Golnas, "Owner/Operator Perspective on Reliability Customer Needs and Field Data," 2011 (cit. on p. 7).

[33] D. C. Jordan and S. R. Kurtz, "Photovoltaic Degradation Rates-an Analytical Review," *Progress in Photovoltaics: Research and Applications*, vol. 21, no. 1, pp. 12–29, Jan. 2013. DOI: 10.1002/pip.1182 (cit. on p. 7).

[34] ——, "Field Performance of 1.7 GW of Photovoltaic Systems," *IEEE Journal of Photovoltaics*, vol. 5, no. 1, pp. 243–249, Jan. 2015. DOI: 10.1109/JPHOTOV.2014.2361667 (cit. on p. 7).

[35] IEC 61215-1:2016, *Terrestrial Photovoltaic (PV) Modules – Design Qualification and Type Approval – Part 1: Test Requirements*, 1st ed., IEC, Ed. Geneva, Switzerland: IEC, 2016, ISBN: 978-2-8322-3206-4 (cit. on pp. 7, 92).

[36] IEC 61724:1998, *Photovoltaic System Performance Monitoring - Guidelines for Measurement, Data Exchange and Analysis*, 1st. Geneva, Switzerland: IEC, 1998 (cit. on pp. 8, 10, 23, 24, 43, 49, 74).

[37] A. Woyte, M. Richter, D. Moser, M. Green, S. Mau, and H. G. Beyer, "Analytical Monitoring of Grid-connected Photovoltaic Systems," International Energy Agency, Report IEA-PVPS T13-03:2014, 2014, p. 90 (cit. on p. 8).

[38] A. Woyte, M. Richter, D. Moser, S. Mau, N. H. Reich, and U. Jahn, "Monitoring of Photovoltaic Systems: Good Practices and Systematic Analyes," in *28th EU-PVSEC*, 2013, pp. 3283–3287. DOI: 10.4229/25thEUPVSEC2010-3AV.2.46 (cit. on p. 8).

[39] D. L. King, "More "Efficient" Methods for Specifying and Monitoring PV System Performance," in *37th IEEE PVSC*, 2011 (cit. on p. 8).

[40] I. Martínez-Marchena, L. Mora-López, M. Piliougine Rocha, and M. Sidrach de Cardona, "An Integrated Software for Monitoring and Evaluation Solar Photovoltaic Installations," in *25th EU-PVSEC*, 2010, pp. 4729–4732. DOI: 10.4229/25thEUPVSEC2010-4BV.1.99 (cit. on p. 8).

[41] A. Phinikarides, C. Shimitra, R. Bourgeon, I. Koumparou, G. Makrides, and G. E. Georghiou, "Development of a Novel Web Application for Automatic Photovoltaic System Performance Analysis and Fault Identification," in *43rd IEEE PVSC*, Portland, OR, USA, 2016, pp. 1736–1740. DOI: 10.1109/PVSC.2016.7749921 (cit. on pp. 8, 129).

[42] Y. Hu, V. Y. Gunapati, P. Zhao, D. Gordon, N. R. Wheeler, M. A. Hossain, T. J. Peshek, L. S. Bruckman, G. Q. Zhang, and R. H. French, "A Nonrelational Data Warehouse for the Analysis of Field and Laboratory Data From Multiple Heterogeneous Photovoltaic Test Sites," *IEEE Journal of Photovoltaics*, vol. PP, no. 99, pp. 1–7, 2016. DOI: 10.1109/JPHOTOV.2016.2626919 (cit. on p. 8).

[43] C. Kopacz, S. Spataru, D. Sera, and T. Kerekes, "Remote and centralized monitoring of PV power plants," IEEE, May 2014, pp. 721–728, ISBN: 978-1-4799-5183-3. DOI: 10.1109/OPTIM.2014.6851005 (cit. on p. 8).

[44] L. Fanni, M. Giussani, M. Marzoli, and M. Nikolaeva-Dimitrova, "How accurate is a commercial monitoring system for photovoltaic plant?" *Progress in Photovoltaics: Research and Applications*, vol. 22, no. 8, pp. 910–922, Aug. 2014. DOI: 10.1002/pip.2328 (cit. on p. 8).

[45] J. Han, I. Lee, and S.-H. Kim, "User-friendly monitoring system for residential PV system based on low-cost power line communication," *IEEE Transactions on Consumer Electronics*, vol. 61, no. 2, pp. 175–180, May 2015. DOI: 10.1109/TCE.2015.7150571 (cit. on p. 8).

[46] S. Spataru, D. Sera, T. Kerekes, and R. Teodorescu, "Photovoltaic array condition monitoring based on online regression of performance model," IEEE, Jun. 2013, pp. 0815–0820. DOI: 10.1109/PVSC.2013.6744271 (cit. on p. 8).

[47] I. Moreno-Garcia, E. Palacios-Garcia, V. Pallares-Lopez, I. Santiago, M. Gonzalez-Redondo, M. Varo-Martinez, and R. Real-Calvo, "Real-Time Monitoring System for a Utility-Scale Photovoltaic Power Plant," *Sensors*, vol. 16, no. 6, p. 770, May 26, 2016. DOI: 10.3390/s16060770 (cit. on p. 8).

[48] A. Drews, A. C. de Keizer, H. G. Beyer, E. Lorenz, J. Betcke, W. G. J. H. M. van Sark, W. Heydenreich, E. Wiemken, S. Stettler, P. Toggweiler, S. Bofinger, M. Schneider, G. Heilscher, and D. Heinemann, "Monitoring and remote failure detection of grid-connected PV systems based on satellite observations," *Solar Energy*, vol. 81, no. 4, pp. 548–564, 2007. DOI: 10.1016/j.solener.2006.06.019 (cit. on p. 8).

[49] A. Kimber, T. Dierauf, L. Mitchell, C. Whitaker, T. U. Townsend, J. Newmiller, D. L. King, J. E. Granata, K. Emery, C. R. Osterwald, D. R. Myers, B. Marion, A. Pligavko, A. F. Panchula, T. Levitsky, J. Forbess, and F. Talmud, "Improved test method to verify the power rating of a photovoltaic (PV) project," in *34th Annual Conference of IEEE Industrial Electronics, IECON*, Jun. 2009, pp. 316–321, ISBN: 978-1-4244-2949-3. DOI: 10.1109/PVSC.2009.5411670 (cit. on p. 8).

[50] D. L. King, J. A. Kratochvil, and W. E. Boyson, "Photovoltaic array performance model," Sandia National Laboratories, SAND2004-3535, 2004 (cit. on p. 8).

[51] S. Ransome, "A review of kWh/kWp measurements, analysis and modelling," in *23rd EU-PVSEC*, vol. 44, 2008, pp. 2795–2800 (cit. on pp. 8, 10).

[52] D. C. Jordan, "Methods for Analysis of Outdoor Performance Data," in *Photovoltaic Module Reliability Workshop*, 2011 (cit. on p. 8).

[53] D. C. Jordan, J. H. Wohlgemuth, and S. R. Kurtz, "Technology and Climate Trends in PV Module Degradation," in *27th EU-PVSEC*, 2012, pp. 3118–3124. DOI: 10.4229/27thEUPVSEC2012-4DO.5.1 (cit. on p. 8).

[54] S. Spataru, D. Sera, T. Kerekes, and R. Teodorescu, "Diagnostic method for photovoltaic systems based on light I–V measurements," *Solar Energy*, vol. 119, pp. 29–44, Sep. 2015. DOI: 10.1016/j.solener.2015.06.020 (cit. on pp. 9, 11).

[55] D. Sera, R. Teodorescu, and P. Rodriguez, "Photovoltaic module diagnostics by series resistance monitoring and temperature and rated power estimation," in *34th Annual Conference of IEEE Industrial Electronics, IECON*, 2009, pp. 2195–2199 (cit. on pp. 9, 11).

[56] D. Sera, "Series Resistance Monitoring for Photovoltaic Modules in the Vicinity of MPP," in *25th EU-PVSEC*, 2010, pp. 4506–4510. DOI: 10.4229/25thEUPVSEC2010-4BV.1.38 (cit. on p. 9).

[57] A. Phinikarides, G. Makrides, and G. E. Georghiou, "Initial performance degradation of an a-Si/a-Si tandem PV array," in *27th EU-PVSEC*, Frankfurt, Germany, 2012, pp. 3267–3270. DOI: 10.4229/27thEUPVSEC2012-4BV.2.16 (cit. on pp. 9, 12, 16, 87).

[58] F. De Lia, S. Castello, and L. Abenante, "Efficiency degradation of c-silicon photovoltaic modules after 22-year continuous field exposure," in *3rd World Conference on Photovoltaic Energy Conversion*, 2003, pp. 2105–2108 (cit. on pp. 9, 17).

[59] Y. Hishikawa and Y. Tsuno, "Calculation Formula for Irradiance and Temperature Correction of the I-V Curves of Solar Cells and Modules by Linear Interpolation/Extrapolation," in *24th EU-PVSEC*, vol. 1, 2009, pp. 3548–3552. DOI: 10.4229/24thEUPVSEC2009-4AV.3.71 (cit. on p. 9).

[60] IEC 61853:2010, *Photovoltaic (PV) Module Performance Testing and Energy Rating - Part 1: Irradiance and Temperature Performance Measurements and Power Rating*, 1st. Geneva, Switzerland: IEC, 2010 (cit. on p. 9).

[61] C. Whitaker, T. U. Townsend, J. Newmiller, D. L. King, W. E. Boyson, J. A. Kratochvil, D. E. Collier, and D. R. Osborn, "Application and validation of a new PV performance characterization method," in *26th IEEE PVSC*, 1997, pp. 1253–1256, ISBN: 0-7803-3767-0. DOI: 10.1109/PVSC.1997.654315 (cit. on p. 9).

[62] S. Smith, T. Townsend, C. Whitaker, and S. Hester, "Photovoltaics for utility-scale applications: Project overview and data analysis," *Solar Cells*, vol. 27, pp. 259–266, 1–4 Oct. 1989. DOI: 10.1016/0379-6787(89)90034-3 (cit. on p. 9).

[63] S. Ransome and J. H. Wohlgemuth, "Predicting kWh/kWp performance for amorphous silicon thin film modules," in *28th IEEE PVSC*, 2000, pp. 1505–1508, ISBN: 0-7803-5772-8. DOI: 10.1109/PVSC.2000.916180 (cit. on p. 9).

[64] J. Adelstein and W. Sekulic, "Small PV Systems Performance Evaluation at NREL' s Outdoor Test Facility Using the PVUSA Power Rating Method," in *DOE Solar Energy Technologies Program Review Meeting*, United States. Dept. of Energy, 2005 (cit. on p. 9).

[65] D. C. Jordan and S. R. Kurtz, "The Dark Horse of Evaluating Long-Term Field Performance—Data Filtering," *IEEE Journal of Photovoltaics*, vol. 4, no. 1, pp. 317–323, Jan. 2014. DOI: 10.1109/JPHOTOV.2013.2282741 (cit. on pp. 10, 11, 16, 44).

[66] S. Ransome, "The Present Status of kWh/kWp Measurements and Modelling," in *25th EU-PVSEC*, 2010, pp. 3873–3878. DOI: 10.4229/25thEUPVSEC2010-4CO.20.3 (cit. on p. 10).

[67] B. Zinsser, G. Makrides, W. Schmitt, G. E. Georghiou, and J. H. Werner, "Annual Energy Yield of 13 Photovoltaic Technologies in Germany and in Cyprus," in *22nd EU-PVSEC*, 2007, pp. 3114–3117 (cit. on p. 10).

[68] B. Marion, J. Adelstein, K. Boyle, H. Hayden, B. Hammond, T. Fletcher, B. Canada, D. Narang, A. Kimber, L. Mitchell, G. Rich, and T. U. Townsend, "Performance parameters for grid-connected PV systems," in *31st IEEE PVSC*, 2005, pp. 1601–1606, ISBN: 0-7803-8707-4. DOI: 10.1109/PVSC.2005.1488451 (cit. on p. 10).

[69] G. Makrides, B. Zinsser, G. E. Georghiou, M. Schubert, and J. H. Werner, "Degradation of different photovoltaic technologies under field conditions," in *35th IEEE PVSC*, 2010, pp. 2332–2337 (cit. on pp. 10, 17).

[70] A. Phinikarides, G. Makrides, and G. E. Georghiou, "Comparison of analysis methods for the calculation of degradation rates of different photovoltaic technologies," in *28th EU-PVSEC*, Paris, France, 2013, pp. 3973–3976. DOI: 10.4229/28thEUPVSEC2013-5BV.4.39 (cit. on pp. 10, 14, 74).

[71] D. L. King, J. A. Kratochvil, and W. E. Boyson, "Temperature coefficients for PV modules and arrays: Measurement methods, difficulties, and results," in *26th IEEE PVSC*, 1997, pp. 1183–1186, ISBN: 0-7803-3767-0. DOI: 10.1109/PVSC.1997.654300 (cit. on pp. 10, 42).

[72] E. Skoplaki and J. A. Palyvos, "On the temperature dependence of photovoltaic module electrical performance: A review of efficiency/power correlations," *Solar Energy*, vol. 83, no. 5, pp. 614–624, May 2009. DOI: 10.1016/j.solener.2008.10.008 (cit. on p. 10).

[73] A. Virtuani, D. Strepparava, and G. Friesen, "A simple approach to model the performance of photovoltaic solar modules in operation," *Solar Energy*, vol. 120, pp. 439–449, 2015. DOI: 10.1016/j.solener.2015.07.045 (cit. on p. 10).

[74] S. Silvestre, M. A. D. Silva, A. Chouder, D. Guasch, and E. Karatepe, "New procedure for fault detection in grid connected PV systems based on the evaluation of current and voltage indicators," *Energy Conversion and Management*, vol. 86, pp. 241–249, 2014. DOI: 10.1016/j.enconman.2014.05.008 (cit. on p. 11).

[75] A. Chouder and S. Silvestre, "Automatic supervision and fault detection of PV systems based on power losses analysis," *Energy Conversion and Management*, vol. 51, no. 10, pp. 1929–1937, 2010. DOI: 10.1016/j.enconman.2010.02.025 (cit. on p. 11).

[76] S. Silvestre, A. Chouder, and E. Karatepe, "Automatic fault detection in grid connected PV systems," *Solar Energy*, vol. 94, pp. 119–127, 2013. DOI: 10.1016/j.solener.2013.05.001 (cit. on pp. 11, 129).

[77] R. Khenfer, M. Mostefai, S. Benahdouga, and M. Maddad, "Faults Detection in a Photovoltaic Generator by Using Matlab Simulink and the chipKIT Max32 Board," *International Journal of Photoenergy*, vol. 2014, pp. 1–9, 2014. DOI: 10.1155/2014/350345 (cit. on p. 11).

[78] J. Zhu, Y. Qiu, T. R. Betts, and R. Gottschalg, "Outlier identification in outdoor measurement data - effects of different strategies on the performance descriptors of photovoltaic modules," in *34th IEEE PVSC*, 2009, pp. 2–6, ISBN: 978-1-4244-2949-3. DOI: 10.1109/PVSC.2009.5411160 (cit. on p. 11).

[79] J. Hedstrom and L. Palmblad, "Performance of old PV modules - Measurement of 25 years old crystalline silicon modules," 06:71, 2006 (cit. on p. 11).

[80] T. Carlsson, K. Astrom, P. Konttinen, and P. Lund, "Data filtering methods for determining of performance parameters in photovoltaic module field tests," *Progress in Photovoltaics: Research and Applications*, vol. 14, no. 4, pp. 329–340, 2006 (cit. on pp. 11, 44).

[81] W. Herrmann, A. Steland, and W. Herff, "Sampling Procedures for the Validation of PV Module Output Power Specification," in *24th EU-PVSEC*, 2009, pp. 3540–3547. DOI: 10.4229/24thEUPVSEC2009-4AV.3.70 (cit. on p. 11).

[82] C. C. Aggarwal, *Outlier Analysis*. New York, NY: Springer New York, 2013, ISBN: 978-1-4614-6395-5 (cit. on pp. 11, 56).

[83] M. Daszykowski, K. Kaczmarek, Y. Vander Heyden, and B. Walczak, "Robust statistics in data analysis - A review. Basic concepts," *Chemometrics and Intelligent Laboratory Systems*, vol. 85, no. 2, pp. 203–219, 2007. DOI: 10.1016/j.chemolab.2006.06.016 (cit. on pp. 11, 44).

[84] P. J. Rousseeuw and M. Hubert, "Robust statistics for outlier detection," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 1, no. 1, pp. 73–79, Jan. 2011. DOI: 10.1002/widm.2 (cit. on p. 11).

[85] Y. Zhao, B. Lehman, R. Ball, J. Mosesian, and J.-F. de Palma, "Outlier detection rules for fault detection in solar photovoltaic arrays," in *28th IEEE Applied Power Electronics Conference and Exposition (APEC)*, 2013, pp. 2913–2920. DOI: 10.1109/APEC.2013.6520712 (cit. on p. 11).

[86] Y. Zhao, L. Yang, B. Lehman, J.-F. de Palma, J. Mosesian, and R. Lyons, "Decision tree-based fault detection and classification in solar photovoltaic arrays," IEEE, Feb. 2012, pp. 93–99. DOI: 10.1109/APEC.2012.6165803 (cit. on p. 11).

[87] Y. Zhao, "Fault Detection, Classification and Protection in Solar Photovoltaic Arrays," PhD Thesis, Northeastern University, Boston, Massachusetts, 2015, 173 pp. (cit. on p. 11).

[88] P. Casas, S. Vaton, L. Fillatre, and I. Nikiforov, "Optimal volume anomaly detection and isolation in large-scale IP networks using coarse-grained measurements," *Computer Networks*, vol. 54, no. 11, pp. 1750–1766, Aug. 2010. DOI: 10.1016/j.comnet.2010.01.013 (cit. on pp. 11, 12).

[89] D.-S. Pham, S. Venkatesh, M. Lazarescu, and S. Budhaditya, "Anomaly detection in large-scale data stream networks," *Data Mining and Knowledge Discovery*, vol. 28, no. 1, pp. 145–189, Jan. 2014. DOI: 10.1007/s10618-012-0297-3 (cit. on p. 12).

[90] A. Lakhina, M. Crovella, and C. Diot, "Diagnosing network-wide traffic anomalies," in *ACM SIGCOMM Computer Communication Review*, vol. 34, ACM, 2004, pp. 219–230. DOI: 10.1145/1015467.1015492 (cit. on p. 12).

[91] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *Journal of the ACM*, vol. 58, no. 3, pp. 1–37, May 2011. DOI: 10.1145/1970392.1970395. pmid: 19686071 (cit. on pp. 12, 44, 63).

[92] A. Artemov and E. Burnaev, "Detecting Performance Degradation of Software-Intensive Systems in the Presence of Trends and Long-Range Dependence," *arXiv preprint arXiv:1609.07662*, 2016 (cit. on p. 12).

[93] The Edison Electric Institute (EEI), *Uniform Business Practices for Unbundled Electricity Metering*. 2000, vol. 2 (cit. on p. 12).

[94] J. L. Schafer and J. W. Graham, "Missing data: Our view of the state of the art.," *Psychological Methods*, vol. 7, no. 2, pp. 147–177, 2002. DOI: 10.1037//1082-989X.7.2.147 (cit. on pp. 12, 67).

[95] K.-J. Hsu, "Missing Data Interpolation of Power Generation for Photovoltaic System," in *Sustainability in Energy and Buildings*, R. J. Howlett, L. C. Jain, and S. H. Lee, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 79–87. DOI: 10.1007/978-3-642-03454-1_9 (cit. on p. 12).

[96] S. Ransome, "Array Performance Analysis Using Imperfect or Incomplete Input Data," in *23rd EU-PVSEC*, vol. 44, 2008, pp. 1–5 (cit. on p. 12).

[97] E. Koubli, D. Palmer, P. Rowley, and R. Gottschalg, "Inference of missing data in photovoltaic monitoring datasets," *IET Renewable Power Generation*, pp. 1–6, Jan. 2016. DOI: 10.1049/iet-rpg.2015.0355 (cit. on p. 12).

[98] J. Honaker and G. King, "What to Do about Missing Values in Time-Series Cross-Section Data," *American Journal of Political Science*, vol. 54, no. 2, pp. 561–581, Apr. 2010. DOI: 10.1111/j.1540-5907.2010.00447.x (cit. on pp. 12, 13, 69).

[99] T. Hastie, R. Tibshirani, and G. Sherlock, "Imputing missing data for gene expression arrays," Division of Biostatistics, Stanford University, 1999, pp. 1–9 (cit. on p. 12).

[100] J. Peppanen, X. Zhang, S. Grijalva, and M. J. Reno, "Handling Bad or Missing Smart Meter Data through Advanced Data Imputation," in *IEEE PES Innovative Smart Grid Technologies (ISGT)*, 2016 (cit. on p. 12).

[101] N. Horton and K. P. Kleinman, "Much Ado About Nothing: A Comparison of Missing Data Methods and Software to Fit Incomplete Data Regression Models," *The American Statistician*, vol. 61, pp. 79–90, 2007 (cit. on p. 13).

[102] M. Blackwell, J. Honaker, and G. King, "A Unified Approach to Measurement Error and Missing Data: Overview and Applications," *Sociological Methods & Research*, forthcoming, Jun. 2015. DOI: 10.1177/0049124115585360 (cit. on p. 13).

[103] A. Phinikarides, G. Makrides, N. Kindyni, and G. E. Georghiou, "Comparison of trend extraction methods for calculating performance loss rates of different photovoltaic technologies," in *40th IEEE PVSC*, Denver, CO, 2014, pp. 3211–3215. DOI: 10.1109/PVSC.2014.6925619 (cit. on pp. 13, 17, 44, 74).

[104] D. C. Jordan and S. R. Kurtz, "Analytical improvements in PV degradation rate determination," in *35th IEEE PVSC*, Jun. 2010, pp. 2688–2693, ISBN: 978-1-4244-5890-5. DOI: 10.1109/PVSC.2010.5617074 (cit. on pp. 13–17).

[105] D. C. Jordan and S. R. Kurtz, "Thin-film reliability trends toward improved stability," in *37th IEEE PVSC*, 2011, pp. 827–832, ISBN: 978-1-4244-9965-6. DOI: 10.1109/PVSC.2011.6186081 (cit. on p. 13).

[106] R. Romero, M. Bennett, and L. Bilella, "Review of SunPower Fleet-Wide System Degradation Study using Year-over-Year Performance Index Analysis," Black & Veatch, B&V Project Number 177395, 2012 (cit. on pp. 13, 17).

[107] H. Theil, "A Rank-Invariant Method of Linear and Polynomial Regression Analysis," in *Henri Theil's Contributions to Economics and Econometrics*, B. Raj and J. Koerts, Eds., vol. 23, Dordrecht: Springer Netherlands, 1992, pp. 345–381, ISBN: 978-94-010-5124-8 (cit. on p. 14).

[108] P. K. Sen, M. L. Puri, and P. R. Krishnaiah, "On robust nonparametric estimation in some multivariate linear models," in *2nd International Symposium on Multivariate Analysis*, 1968, pp. 33–52 (cit. on p. 14).

[109] R. R. Wilcox, *Fundamentals of Modern Statistical Methods: Substantially Improving Power and Accuracy*. New York: Springer, 2001, 258 pp., ISBN: 978-0-387-95157-7 (cit. on p. 14).

[110] D. M. Miller and D. Williams, "Shrinkage estimators of time series seasonal factors and their effect on forecasting accuracy," *International Journal of Forecasting*, vol. 19, no. 4, pp. 669–684, Oct. 2003. DOI: 10.1016/S0169-2070(02)00077-8 (cit. on p. 14).

[111] S. G. Makridakis, S. Wheelwright, and R. J. Hyndman, *Forecasting: Methods and Applications*. Wiley, 1998 (cit. on p. 14).

[112] A. N. Dunea, D. N. Dunea, V. I. Moise, and M. F. Olariu, "Forecasting methods used for performance's simulation and optimization of photovoltaic grids," in *IEEE Porto Power Tech Proceedings*, IEEE, 2001, p. 5, ISBN: 0-7803-7139-9. DOI: 10.1109/PTC.2001.964864 (cit. on p. 14).

[113] G. E. P. Box and G. M. Jenkins, "Some recent advances in forecasting and control," *Applied Statistics*, vol. 23, no. 2, pp. 158–179, 1968. JSTOR: 10.2307/2985674 (cit. on pp. 15, 77).

[114] G. E. P. Box, G. M. Jenkins, and G. C. Reinsel, *Time Series Analysis: Forecasting and Control*. Prentice Hall, 1976, ISBN: 978-0-13-060774-4 (cit. on p. 15).

[115] D. L. Staebler, R. S. Crandall, and S. R. Williams, "Stability of n-i-p amorphous silicon solar cells," *Applied Physics Letters*, vol. 39, no. 9, pp. 733–735, 1981 (cit. on p. 16).

[116] G. Makrides, B. Zinsser, M. Schubert, and G. E. Georghiou, "Performance loss rate of twelve photovoltaic technologies under field conditions using statistical techniques," *Solar Energy*, vol. 103, pp. 28–42, May May 2014. DOI: 10.1016/j.solener.2014.02.011 (cit. on p. 16).

[117] D. C. Jordan, R. Smith, C. R. Osterwald, E. Gelak, and S. R. Kurtz, "Outdoor PV degradation comparison," in *35th IEEE PVSC*, Jun. 2010, pp. 2694–2697, ISBN: 978-1-4244-5890-5. DOI: 10.1109/PVSC.2010.5616925 (cit. on p. 16).

[118] J. E. Granata, W. E. Boyson, J. A. Kratochvil, and M. Quintana, "Long-term performance and reliability assessment of 8 PV arrays at Sandia National Laboratories," in *34th IEEE PVSC*, Jun. 2009, pp. 1486–1491, ISBN: 978-1-4244-2949-3. DOI: 10.1109/PVSC.2009.5411336 (cit. on p. 16).

[119] M. B. Strobel, T. R. Betts, G. Friesen, H. G. Beyer, and R. Gottschalg, "Uncertainty in Photovoltaic performance parameters – dependence on location and material," *Solar Energy Materials and Solar Cells*, vol. 93, pp. 1124–1128, 6-7 Jun. 2009. DOI: 10.1016/j.solmat.2009.02.003 (cit. on p. 16).

[120] M. Vazquez and I. Rey-Stolle, "Photovoltaic module reliability model based on field degradation studies," *Progress in Photovoltaics: Research and Applications*, vol. 16, no. 5, pp. 419–433, 2008. DOI: 10.1002/pip (cit. on p. 16).

[121] R. Rüther and L. Nascimento, "Long-term performance of the first grid-connected, building-integrated amorphous silicon PV installation in Brazil," in *35th IEEE PVSC*, 2010, pp. 2283–2286, ISBN: 978-1-4244-5892-9 (cit. on p. 17).

[122] G.-H. Kang, K.-S. Kim, H.-E. Song, G.-J. Yu, H.-K. Ahn, and D.-Y. Han, "Investigation of Aging Phenomenon and Power Drop Rate with Field Exposed PV Modules," in *25th EU-PVSEC*, 2010, pp. 4015–4018. DOI: 10.4229/25thEUPVSEC2010-4AV.3.17 (cit. on p. 17).

[123] A. Realini, E. Burà, N. Cereghetti, D. Chianese, and S. Rezzonico, "Mean time before failure of photovoltaic modules (MTBF-PVm)," Federal Office for Education and Science, BBW 99.0579, 2003 (cit. on p. 17).

[124] R. Smith, D. C. Jordan, and S. R. Kurtz, "Outdoor PV Module Degradation of Current-Voltage Parameters," in *World Renewable Energy Forum*, 2012 (cit. on p. 17).

[125] K. J. Sauer, "Real-world challenges and opportunities in degradation rate analysis for commercial PV systems," in *37th IEEE PVSC*, 2011, pp. 3208–3212, ISBN: 978-1-4244-9965-6 (cit. on p. 17).

[126] M. Anderson, Z. Defreitas, and E. F. Hasselbrink, "A System Degradation Study of 445 Systems Using Year-over-Year Performance Index Analysis," 2012 (cit. on p. 17).

[127] A. Phinikarides, G. Makrides, and G. E. Georghiou, "A comprehensive methodology for outdoor and indoor degradation studies on photovoltaic modules," in *3rd International Conference on Renewable Energy Sources & Energy Efficiency*, Nicosia, Cyprus, 2011, pp. 85–93 (cit. on p. 20).

[128] JCGM 100:2008, *GUM 1995 with Minor Corrections: Evaluation of Measurement Data - Guide to the Expression of Uncertainty in Measurement*, 1st. JCGM, 2008 (cit. on pp. 22, 80, 93).

[129] C. E. Shannon, "A Mathematical Theory of Communication," *Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, Jul. 1948. DOI: 10.1002/j.1538-7305.1948.tb01338.x (cit. on p. 23).

[130] ——, "Communication In The Presence Of Noise," *Proceedings of the IRE*, vol. 37, pp. 10–21, 1949 (cit. on p. 23).

[131] A. Driesse, J. S. Stein, D. M. Riley, and C. Carmignani, "Sampling and Filtering in Photovoltaic System Performance Monitoring," Sandia National Laboratories, SAND2014 - 19137, Oct. 2014 (cit. on p. 23).

[132] C. Schuss, B. Eichberger, and T. Rahkonen, "Impact of sampling interval on the accuracy of estimating the amount of solar energy," in *IEEE International Instrumentation and Measurement Technology Conference Proceedings (I2MTC)*, IEEE, 2016, pp. 1–6. DOI: 10.1109/I2MTC.2016.7520566 (cit. on p. 23).

[133] G. Reikard, "Predicting solar radiation at high resolutions: A comparison of time series forecasts," *Solar Energy*, vol. 83, no. 3, pp. 342–349, Mar. 2009. DOI: 10.1016/j.solener.2008.08.007 (cit. on p. 24).

[134] M. Zamo, O. Mestre, P. Arbogast, and O. Pannekoucke, "A benchmark of statistical regression methods for short-term forecasting of photovoltaic electricity production, Part I: Deterministic forecast of hourly production," *Solar Energy*, vol. 105, pp. 792–803, 2014. DOI: 10.1016/j.solener.2013.12.006 (cit. on p. 24).

[135] H. G. Beyer, "Handling of Small Scale Structures of the Irradiance Field for Solar Energy System Analysis – A Review," *Energy Procedia*, vol. 97, pp. 141–148, Nov. 2016. DOI: 10.1016/j.egypro.2016.10.039 (cit. on p. 24).

[136] B. Amrouche, A. Guessoum, and M. Belhamel, "A simple behavioural model for solar module electric characteristics based on the first order system step response for MPPT study and comparison," *Applied Energy*, vol. 91, no. 1, pp. 395–404, Mar. 2012. DOI: 10.1016/j.apenergy.2011.09.036 (cit. on p. 24).

[137] Digital Signal Processing Committe, *Programs for Digital Signal Processing*. Piscataway, NJ, USA: IEEE Press, 1979, ISBN: 0-87942-127-4 (cit. on p. 24).

[138] J. Kaiser and M. Dolan, "An update on "Programs for digital signal processing"," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, no. 5, pp. 1102–1105, Oct. 1981. DOI: 10.1109/TASSP.1981.1163670 (cit. on p. 24).

[139] IEC 60904-9:2007, *Photovoltaic Devices - Part 9: Solar Simulator Performance Requirements*, 2nd. Geneva, Switzerland: IEC, 2007 (cit. on pp. 31, 87).

[140] J. J. Michalsky, "The Astronomical Almanac's algorithm for approximate solar position (1950–2050)," *Solar Energy*, vol. 40, no. 3, pp. 227–235, 1988. DOI: 10.1016/0038-092X(88)90045-X (cit. on pp. 33, 51).

[141] IEC 61215-1-2:2016, *Terrestrial Photovoltaic (PV) Modules – Design Qualification and Type Approval – Part 1-2: Test Procedures*, 1st ed., IEC, Ed. Geneva, Switzerland: IEC, 2016, ISBN: 978-2-8322-3205-7 (cit. on pp. 33, 87).

[142] IEC 61646:2007, *Thin-Film Terrestrial Photovoltaic (PV) Modules - Design Qualification and Type Approval*, 2nd. Geneva, Switzerland: IEC, 2007 (cit. on p. 33).

[143] J. Polo, W. Fernandez-Neira, and M. Alonso-García, "On the use of reference modules as irradiance sensor for monitoring and modelling rooftop PV systems," *Renewable Energy*, vol. 106, pp. 186–191, Jun. 2017. DOI: 10.1016/j.renene.2017.01.026 (cit. on p. 35).

[144] S. S. Shapiro and M. B. Wilk, "An analysis of variance test for normality (complete samples)," *Biometrika*, vol. 52, pp. 591–611, 3-4 Dec. 1965. DOI: 10.1093/biomet/52.3-4.591 (cit. on p. 39).

[145] A. Kolmogorov, "Sulla Determinazione Empirica di una Legge di Distribuzione," *Giornale dell'Istituto Italiano degli Attuari*, vol. 4, pp. 83–91, 1933 (cit. on p. 39).

[146] N. Smirnov, "Estimate of Deviation Between Empirical Distribution Functions in Two Independent Samples," *Bulletin Moscow University*, vol. 2, no. 2, pp. 3–16, 1939 (cit. on p. 39).

[147] W. J. Conover, "Distribution-free methods in statistics," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 1, no. 2, pp. 199–207, 2009. DOI: 10.1002/wics.28 (cit. on p. 39).

[148] T. Elwood, W. Bennett, T. Lai, and K. Simmons-Potter, "In-situ comparison of thermal measurement technologies for interpretation of PV module temperature derating effects," in *Proc. SPIE 9938, Reliability of Photovoltaic Cells, Modules, Components, and Systems IX*, N. G. Dhere, J. H. Wohlgemuth, and K. Sakurai, Eds., Sep. 26, 2016, 99380Q. DOI: 10.1117/12.2237934 (cit. on p. 42).

[149] B. Mihaylov, T. R. Betts, A. Pozza, H. Mullejans, and R. Gottschalg, "Uncertainty Estimation of Temperature Coefficient Measurements of PV Modules," *IEEE Journal of Photovoltaics*, vol. 6, no. 6, pp. 1554–1563, Nov. 2016. DOI: 10.1109/JPHOTOV.2016.2598259 (cit. on p. 42).

[150] E. Skoplaki and J. A. Palyvos, "Operating temperature of photovoltaic modules: A survey of pertinent correlations," *Renewable Energy*, vol. 34, no. 1, pp. 23–29, Jan. 2009. DOI: 10.1016/j.renene.2008.04.009 (cit. on p. 42).

[151] G. Makrides, B. Zinsser, G. E. Georghiou, M. Schubert, and J. H. Werner, "Temperature behaviour of different photovoltaic systems installed in Cyprus and Germany," *Solar Energy Materials and Solar Cells*, vol. 93, pp. 1095–1099, 6-7 Jun. 2009. DOI: 10.1016/j.solmat.2008.12.024 (cit. on p. 42).

[152] V. Gómez, A. Maravall, and D. Peña, "Missing observations in ARIMA models: Skipping approach versus additive outlier approach," *Journal of Econometrics*, vol. 88, no. 2, pp. 341–363, Feb. 1999. DOI: 10.1016/S0304-4076(98)00036-0 (cit. on p. 44).

[153] D. Evagorou, a. Kyprianou, P. L. Lewin, a. Stavrou, V. Efthymiou, a. C. Metaxas, and G. E. Georghiou, "Feature extraction of partial discharge signals using the wavelet packet transform and classification with a probabilistic neural network," *IET Science, Measurement & Technology*, vol. 4, no. 3, p. 177, 2010. DOI: 10.1049/iet-smt.2009.0023 (cit. on p. 44).

[154] R. J. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice*. OTexts, 2012 (cit. on p. 44).

[155] J. Y. Ye, T. Reindl, A. G. Aberle, and T. M. Walsh, "Performance Degradation of Various PV Module Technologies in Tropical Singapore," *IEEE Journal of Photovoltaics*, vol. 4, no. 5, pp. 1288–1294, Sep. 2014. DOI: 10.1109/JPHOTOV.2014.2338051 (cit. on p. 44).

[156] E. B. Dagum and C. Dagum, "Stochastic and deterministic trend models," *Statistica*, vol. 66, no. 3, pp. 269–280, 2006 (cit. on p. 44).

[157] H. Wickham, *Advanced R*, 1st. Boca Raton, FL: Chapman and Hall/CRC, Sep. 25, 2014, 476 pp., ISBN: 978-1-4665-8696-3 (cit. on p. 44).

[158] H. Wickham and R. Francois, *Dplyr: A Grammar of Data Manipulation*. 2016 (cit. on p. 44).

[159] H. Wickham, *Purrr: Functional Programming Tools*. 2016 (cit. on p. 44).

[160] A. Kyprianou, A. Phinikarides, G. Makrides, and G. E. Georghiou, "Robust Principal Component Analysis For Computing The Degradation Rates Of Different Photovoltaic Systems," in *29th EU-PVSEC*, Amsterdam, 2014, pp. 2939–2942. DOI: 10.4229/EUPVSEC20142014-5BV.2.41 (cit. on p. 48).

[161] ——, "Definition and Computation of the Degradation Rates of Photovoltaic Systems of Different Technologies With Robust Principal Component Analysis," *IEEE Journal of Photovoltaics*, vol. 5, no. 6, pp. 1698–1705, Nov. 2015. DOI: 10.1109/JPHOTOV.2015.2478065 (cit. on p. 48).

[162] A. Phinikarides, G. Makrides, and G. E. Georghiou, "Estimation of the Degradation Rate of Fielded Photovoltaic Arrays in the Presence of Measurement Outages," in *32nd EU-PVSEC*, Munich, Germany, 2016, pp. 1754–1757, ISBN: 3-936338-41-8. DOI: 10.4229/EUPVSEC20162016-5DO.12.6 (cit. on p. 48).

[163] D. H. Hathaway, "The Solar Cycle," *Living Reviews in Solar Physics*, vol. 12, no. 1, p. 4, Dec. 1, 2015. DOI: 10.1007/lrsp-2015-4 (cit. on p. 49).

[164] J. G. Da Silva Fonseca Junior, T. Oozeki, H. Ohtake, K. ichi Shimose, T. Takashima, and K. Ogimoto, "Regional forecasts and smoothing effect of photovoltaic power generation in Japan: An approach with principal component analysis," *Renewable Energy*, vol. 68, pp. 403–413, 2014. DOI: 10.1016/j.renene.2014.02.018 (cit. on p. 50).

[165] S. Ponce-Alcántara, J. P. Connolly, G. Sánchez, J. M. Míguez, V. Hoffmann, and R. Ordás, "A Statistical Analysis of the Temperature Coefficients of Industrial Silicon Solar Cells," *Energy Procedia*, vol. 55, pp. 578–588, 2014. DOI: 10.1016/j.egypro.2014.08.029 (cit. on p. 50).

[166] C. Berthod, R. Strandberg, G. H. Yordanov, H. G. Beyer, and J. O. Odden, "On the Variability of the Temperature Coefficients of mc-Si Solar Cells with Irradiance," *Energy Procedia*, vol. 92, pp. 2–9, Aug. 2016. DOI: 10.1016/j.egypro.2016.07.002 (cit. on p. 50).

[167] S. Vergura and F. Vacca, "Bootstrap Technique for Analyzing Energy Data from PV Plant," in *International Conference on Clean Electrical Power*, IEEE, 2009, pp. 268–275. DOI: 10.1109/ICCEP.2009.5212046 (cit. on p. 50).

[168] A. Gelman and A. Vehtari, "Comment," *Journal of the American Statistical Association*, vol. 109, no. 507, pp. 1015–1016, Jul. 3, 2014. DOI: 10.1080/01621459.2014.906153 (cit. on p. 50).

[169] I. Stanimirova, B. Walczak, D. L. Massart, and V. Simeonov, "A comparison between two robust PCA algorithms," *Chemometrics and Intelligent Laboratory Systems*, vol. 71, no. 1, pp. 83–95, 2004. DOI: 10.1016/j.chemolab.2003.12.011 (cit. on p. 56).

[170] M. G. Kendall, "A New Measure of Rank Correlation," *Biometrika*, vol. 30, p. 81, 1/2 Jun. 1938. DOI: 10.2307/2332226. JSTOR: 2332226 (cit. on p. 62).

[171] H. B. Mann, "Nonparametric Tests Against Trend," *Econometrica*, vol. 13, no. 3, p. 245, Jul. 1945. DOI: 10.2307/1907187. JSTOR: 1907187 (cit. on p. 62).

[172] Z. Lin, M. Chen, and Y. Ma, "The Augmented Lagrange Multiplier Method for Exact Recovery of Corrupted Low-Rank Matrices," *arXiv preprint arXiv:1009.5055*, 2010. arXiv: 1009.5055 (cit. on p. 63).

[173] N. B. Erichson, S. Voronin, S. L. Brunton, and J. N. Kutz, "Randomized Matrix Decompositions using R," Aug. 6, 2016. arXiv: 1608.02148 [cs, stat] (cit. on pp. 63, 127).

[174] S. Voronin and P.-G. Martinsson, "RSVDPACK: An implementation of randomized algorithms for computing the singular value, interpolative, and CUR decompositions of matrices on multi-core and GPU architectures," Feb. 18, 2015. arXiv: 1502.05366 [cs, math] (cit. on pp. 63, 127).

[175] N. Halko, P. G. Martinsson, and J. A. Tropp, "Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions," *SIAM Review*, vol. 53, no. 2, pp. 217–288, Jan. 2011. DOI: 10.1137/090771806 (cit. on p. 64).

[176] G. T. Klise and J. S. Stein, "Models Used to Assess the Performance of Photovoltaic Systems," Sandia National Laboratories, SAND2009-8258, 2009 (cit. on p. 67).

[177] A. Gelman and J. Hill, *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, 2006, ISBN: 978-0-521-86706-1 (cit. on p. 69).

[178] A. P. A. P. Dempster, N. M. N. M. Laird, and D. B. D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society Series B Methodological*, vol. 39, no. 1, pp. 1–38, 1977. DOI: 10.2307/2984875. pmid: 9501024 (cit. on p. 70).

[179] S. Van Buuren and K. Groothuis-Oudshoorn, "Multivariate Imputation by Chained Equations," *Journal Of Statistical Software*, vol. 45, no. 3, pp. 1–67, 2011. DOI: 10.1177/096228020607446 3. pmid: 22289957 (cit. on p. 70).

[180] C. Turrado, M. López, F. Lasheras, B. Gómez, J. Rollé, and F. Juez, "Missing Data Imputation of Solar Radiation Data under Different Atmospheric Conditions," *Sensors*, vol. 14, no. 11, pp. 20 382–20 399, 2014. DOI: 10.3390/s141120382. pmid: 25356644 (cit. on p. 70).

[181] R Core Team, *R: A Language and Environment for Statistical Computing*, Vienna, Austria, 2015 (cit. on pp. 70, 129).

[182] A. Phinikarides, N. Philippou, G. Makrides, and G. E. Georghiou, "Performance loss rates of different photovoltaic technologies after eight years of operation under warm climate conditions," in *29th EU-PVSEC*, Amsterdam, 2014, pp. 2664–2668. DOI: 10.4229/EUPVSEC20142014-5BV.1.27 (cit. on p. 74).

[183] A. Phinikarides, G. Makrides, and G. E. Georghiou, "Estimation of annual performance loss rates of grid-connected photovoltaic systems using time series analysis and validation through indoor testing at standard test conditions," in *42nd IEEE PVSC*, New Orleans, LA, 2015, pp. 1–5. DOI: 10.1109/PVSC.2015.7355940 (cit. on pp. 74, 87).

[184] A. Phinikarides, G. Makrides, B. Zinsser, M. Schubert, and G. E. Georghiou, "Analysis of photovoltaic system performance time series: Seasonality and performance loss," *Renewable Energy*, vol. 77, pp. 51–63, May 2015. DOI: 10.1016/j.renene.2014.11.091 (cit. on pp. 74, 88).

[185] A. Phinikarides, G. Makrides, N. Kindyni, A. Kyprianou, and G. E. Georghiou, "ARIMA modeling of the performance of different photovoltaic technologies," in *39th IEEE PVSC*, Tampa, FL, Jun. 2013, pp. 797–801, ISBN: 978-1-4799-3299-3. DOI: 10.1109/PVSC.2013.6744268 (cit. on p. 74).

[186] IEC 60891:1992, *Procedures for Temperature and Irradiance Corrections to Measured I-V Characteristics of Crystalline Silicon Photovoltaic Devices (Amendment 1)*, 1st. Geneva, Switzerland: IEC, 1992 (cit. on p. 74).

[187] R. M. Hirsch, J. R. Slack, and R. A. Smith, "Techniques of trend analysis for monthly water quality data," *Water Resources Research*, vol. 18, no. 1, pp. 107–121, Feb. 1982. DOI: 10.1029/WR018i001p00107 (cit. on p. 75).

[188] C. Libiseller and A. Grimvall, "Performance of partial Mann-Kendall tests for trend detection in the presence of covariates," *Environmetrics*, vol. 13, no. 1, pp. 71–84, Feb. 2002. DOI: 10.1002/env.507 (cit. on p. 75).

[189] R. B. Cleveland, W. S. Cleveland, J. E. McRae, and I. Terpenning, "STL: A seasonal-trend decomposition procedure based on Loess," *Journal of Official Statistics*, vol. 6, no. 1, pp. 3–73, 1990 (cit. on p. 76).

[190] W. S. Cleveland, "Robust locally weighted regression and smoothing scatterplots," *Journal of the American Statistical Association*, vol. 74, no. 368, pp. 829–836, 1979. JSTOR: 10.2307/2286407 (cit. on p. 76).

[191] W. S. Cleveland and S. J. Devlin, "Locally weighted regression: An approach to regression analysis by local fitting," *Journal of the American Statistical Association*, vol. 83, no. 403, pp. 596–610, 1988. JSTOR: 10.2307/2289282 (cit. on p. 76).

[192] D. A. Dickey and W. A. Fuller, "Distribution of the estimators for autoregressive time series with a unit root," *Journal of the American Statistical Association*, vol. 74, no. 366, pp. 427–431, 1979 (cit. on p. 77).

[193] D. Kwiatkowski, P. C. Phillips, P. Schmidt, and Y. Shin, "Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root?" *Journal of econometrics*, vol. 54, pp. 159–178, 1-3 1992 (cit. on p. 77).

[194] B. C. Monsell, "The X-13ARIMA-SEATS Seasonal Adjustment Program," in *Proceedings of the Federal Committee on Statistical Methodology Research Conference*, Washington, DC, 2007 (cit. on p. 78).

[195] Deutsche Bundesbank, "The changeover from the seasonal adjustment method Census X-11 to Census X-12-ARIMA," Deutsche Bundesbank, Monthly Report, 1999 (cit. on p. 78).

[196] M. Manna and R. Peronaci, *Seasonal Adjustment*. European Central Bank, 2003, ISBN: ISBN 92-9181-413-X (cit. on p. 78).

[197] J. Shiskin, A. Young, and J. Musgrave, "The X-11 Variant of the Census Method II Seasonal Adjustment Program," Bureau of the Census, US Department of Commerce, 1967 (cit. on p. 78).

[198] E. B. Dagum, *The X-11-ARIMA Seasonal Adjustment Method*. Statistics Canada, 1980, 678–795 (cit. on p. 78).

[199] D. F. Findley, B. C. Monsell, W. R. Bell, M. C. Otto, and B.-C. Chen, "New Capabilities and Methods of the X-12-ARIMA Seasonal-Adjustment Program," *Journal of Business & Economic Statistics*, vol. 16, no. 2, p. 169, Apr. 1998. DOI: 10.2307/1392572. JSTOR: 1392572 (cit. on p. 78).

[200] B. C. Monsell, "Update on the Development of X-13ARIMA-SEATS," in *Proceedings of the Joint Statistical Meetings, Business and Economic Statistics Section*, Alexandria, VA: American Statistical Association, 2009 (cit. on p. 78).

[201] I. H. Chang, G. C. Tiao, and B.-C. Chen, "Estimation of Time Series Parameters in the Presence of Outliers," *Technometrics*, vol. 30, no. 2, p. 193, May 1988. DOI: 10.2307/1270165. JSTOR: 1270165 (cit. on p. 79).

[202] W. R. Bell and S. C. Hillmer, "Modeling Time Series With Calendar Variation," *Journal of the American Statistical Association*, vol. 78, no. 383, p. 526, Sep. 1983. DOI: 10.2307/2288114. JSTOR: 2288114 (cit. on p. 79).

[203] A. N. Pettitt, "A Non-Parametric Approach to the Change-Point Problem," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 28, no. 2, pp. 126–135, 1979. DOI: 10.2307/2346729. JSTOR: 2346729 (cit. on p. 82).

[204] A. Phinikarides, G. Makrides, and G. E. Georghiou, "Analysis of the field performance of a double junction amorphous silicon photovoltaic module and its correlation to standardized testing," in *31st EU-PVSEC*, Hamburg, Germany, 2015, pp. 1992–1996. DOI: 10.4229/EUPVSEC20152015-5AV.6.20 (cit. on pp. 87, 88).

[205] C. R. Wronski and X. Niu, "The Limited Relevance of SWE Dangling Bonds to Degradation in High-Quality a-Si:H Solar Cells," *IEEE Journal of Photovoltaics*, vol. 4, no. 3, pp. 778–784, May 2014. DOI: 10.1109/JPHOTOV.2014.2311498 (cit. on p. 88).

[206] D. Wagner and P. Irsigler, "On the annealing behaviour of the Staebler-Wronski effect in a-Si:H," *Applied Physics A Solids and Surfaces*, vol. 35, no. 1, pp. 9–12, Sep. 1984. DOI: 10.1007/BF00620293 (cit. on p. 88).

[207] D. Riley and J. Johnson, "Photovoltaic prognostics and heath management using learning algorithms," in *38th IEEE PVSC*, 2012, pp. 1535–1539, ISBN: 978-1-4673-0064-3. DOI: 10.1109/PVSC.2012.6317887 (cit. on p. 90).

[208] L. Micheli and M. Muller, "An investigation of the key parameters for predicting PV soiling losses: Key parameters for predicting PV soiling losses," *Progress in Photovoltaics: Research and Applications*, 2017. DOI: 10.1002/pip.2860 (cit. on p. 90).

[209] A. Kimber, L. Mitchell, S. Nogradi, and H. Wenger, "The Effect of Soiling on Large Grid-Connected Photovoltaic Systems in California and the Southwest Region of the United States," in *4th IEEE World Conference on Photovoltaic Energy Conversion*, IEEE, 2006, pp. 2391–2395, ISBN: 978-1-4244-0016-4. DOI: 10.1109/WCPEC.2006.279690 (cit. on p. 90).

[210] M. R. Maghami, H. Hizam, C. Gomes, M. A. Radzi, M. I. Rezadad, and S. Hajighorbani, "Power loss due to soiling on solar panel: A review," *Renewable and Sustainable Energy Reviews*, vol. 59, pp. 1307–1316, Jun. 2016. DOI: 10.1016/j.rser.2016.01.044 (cit. on p. 90).

[211] D. Thevenard and S. Pelland, "Estimating the uncertainty in long-term photovoltaic yield predictions," *Solar Energy*, vol. 91, pp. 432–445, 2013. DOI: 10.1016/j.solener.2011.05.006 (cit. on p. 90).

[212] IEC 60904-7:1998, *Photovoltaic Devices - Part 7: Computation of Spectral Mismatch Error Introduced in the Testing of a Photovoltaic Device*, 2nd. Geneva, Switzerland: IEC, 1998 (cit. on p. 92).

[213] R. Fisher, *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd, 1970, ISBN: 0-05-002170-2 (cit. on p. 93).

[214] R. Ebner, S. Zamini, and G. Újvári, "Defect Analysis in Different Photovoltaic Modules Using Electroluminescence (EL) and Infrared (IR)-Thermography," in *25th EU-PVSEC*, 2010, pp. 333–336. DOI: 10.4229/25thEUPVSEC2010-1DV.2.8 (cit. on p. 95).

[215] T. Bouwmans and E. H. Zahzah, "Robust PCA via Principal Component Pursuit: A review for a comparative evaluation in video surveillance," *Computer Vision and Image Understanding*, vol. 122, pp. 22–34, May 2014. DOI: 10.1016/j.cviu.2013.11.009 (cit. on p. 127).

[216] O. Perpiñán, "solaR : Solar Radiation and Photovoltaic Systems with R," *Journal Of Statistical Software*, vol. 50, no. 9, pp. 1–32, 2012. DOI: 10.18637/jss.v050.i09 (cit. on p. 129).

[217] IEEE Std 754-2008, *IEEE Standard for Floating-Point Arithmetic*. Aug. 2008, 70 pp. (cit. on p. 131).

Appendix A

Computational Expense

A.1 Computational cost

There are several techniques that can be used to reduce the cost of a computationally heavy analysis, such as the one developed in this work. One of them is concerned with the programming style and defines explicit and implicit parallelism to take advantage of all available computing resources. Explicit parallel programming can be employed by the programmer to divide a large task that deals with computing independent realizations into smaller tasks and distribute across multiple central processing unit (CPU) cores. As an example, an expensive `for` loop over large lists of variables contained in this work was replaced by many cheaper partitions of computations which were spread to more than one core, resulting in a drastic reduction of computation time. The partitions were split as equally as possible to the number of CPU cores, to create a uniform load across all cores and minimize the overhead of setting up new threads as the old ones complete their job.

On the other hand, usage of an optimizing compiler which automatically extracts the parallelism inherent to computations can be used to provide implicit parallellism. It is important to note though, that only specific kinds of computations can be converted to implicitly parallelized code by the compiler and that implicit and explicit parallellism operate on different domains. Therefore, one is not a replacement of the other. Implicit parallellism was enabled throughout this work, by compiling and linking `R` with high performance linear algebra libraries (i.e. Basic Linear Algebra Subprograms (BLAS) and the Intel®Math Kernel Library (MKL).) Intel®MKL is a math library of highly optimized and multi-threaded routines which provided accelerated linear algebra routines for vector and matrix operations, high-performance vectorized random number generators (RNGs) for several probability distributions and convolution and correlation routines. More information can be found in Appendix B.2.

Another technique that was used to reduce the computational cost was caching. Data objects carrying the results of expensive computations were cached to disk, once computation was complete. The cached data was recalled on demand, provided that the input data and the function used to compute them had not changed. This was achieved by storing

the hash of the input data and the function text as metadata and linking it to the saved data on disk. If the hashes did not match, the cached result was discarded and had to be recomputed. The concept of caching was very important throughout this work, since past measurement data were static. It did not make sense to apply the same functions more than once, over the same set of data, which is especially desirable in the online monitoring paradigm. Specifically, in estimating statistics over daily or monthly groups, these were computed once, saved and reused.

A.2 Bottlenecks

The major bottlenecks in the proposed data analysis methodology were the ALM method used to optimize PCP for obtaining RPCA, the bootstrap of the boxplot outlier rule and the bootstrap for estimating confidence intervals on the R_{D_E}. All three of these procedures were iterative in nature and suffered from the curse of dimensionality. This was especially important in this work, as the number of PV systems under test (11), combined with the high number of constructed metrics meant that each analysis would take a significant amount of time to complete.

A.2.1 Bootstrap

In the case of the bootstrap, the computation of each realization of the bootstrapped statistic was independent of each other, therefore this presented what is called an embarrassingly parallel problem. Such problems are easily solved by spreading the computational load across many CPU cores, which can either belong to the local machine or a remote cluster.

In `R`, there are two main types of parallel clusters:

1. a socket type cluster, which runs `Rscript` on the specified host(s) or CPUs to set up worker processes which listen on their own socket for expressions to evaluate, and return the results as serialized objects

2. a fork type cluster, which forks the main *R* process and links all workers to the same address space

On both Linux and Windows machines, parallelism can be implemented by creating a socket cluster. A socket cluster has the advantage that it can be started on a remote host and not only on the local host, in contrast to the fork cluster type. On the other hand, a socket cluster operates each worker process in their own memory address space, effectively multiplying the memory requirements for the analysis by the number of parallel workers started.

The fork cluster type is exclusive to Linux, as it can only work with operating systems that support the fork system call. It provided superior performance in comparison to the socket type by avoiding the overhead of setting up separate workers and duplicating the

memory space of the main *R* process for each one. It thus resulted in much less memory consumption, an important fact, given the amount of data produced in this work, which enabled the analysis on commodity hardware without running out of memory. Whereas the minimum memory requirements for running the analysis on Linux were 16GB RAM, on Windows it was at least 24GB. Throughout this work, Linux and the fork cluster type were used extensively, although the analysis was also ported and can thus run on a Windows system.

Regarding application of the bootstrap in this analysis, fixing the bottleneck with explicit parallelization resulted in speedup up to the maximum number of cores used. To put this into perspective, computing the daily bootstrapped boxplot statistics and confidence intervals required 61 s per PV system time series or 18 ms per time series, per day for 1000 bootstrap samples. This was quite fast on its own for a single system and a single day, but quite slow as an absolute measure of computational cost when the whole data set was considered. By spreading computation across the available cores, the time required for a single system and single day was still 18 ms, but the total computation time across all days for a single system was reduced by a factor of 4 on an i7-2600K processor.

A.2.2 Robust Principal Component Analysis

On the other hand, the RPCA could not be treated the same way as the bootstrap, as the algorithm minimizes a cost function, with input from the previous step. Therefore, ways to reduce computation time were either to optimize the convergence of the SVD function, use a different solver [215] or accelerate linear algebra computations by linking with a multi-threaded library. In this work, the second and third options were investigated. Linking with a multi-threaded library is described in Appendix B.2. This enabled multi-threading of the SVD function used to realize RPCA. The cost was reduced by a factor of the number of physical cores in the system. More specifically, decomposition of the $\mathbf{M}_{96\times3287}$ matrix of P_A measurements for a single PV system required 105 seconds and 2300 iterations to converge, using the accelerated linear algebra libraries. In a real world application of monitoring hundreds, even thousands of PV systems, it is evident that this would not be a good candidate for outlier detection since the benefits of RPCA would be negated by its slow execution time.

Using the exact ALM for RPCA was very slow, therefore different solvers were investigated to enable the use of the RPCA for this analysis. A faster alternative was the randomized accelerated IALM which used a randomized implementation of the SVD [173, 174] which was able to obtain the robust separation in significantly less amount of time. To quantify the complexity of different implementations, each algorithm was run iteratively for a minimum of 20 repetitions. The elapsed time was recorded and used to create the plot shown in Fig. A.1, which displays a boxplot with a rotated kernel density plot on each side (otherwise known as a violin plot). In this way, the minimum, maximum and mean time required for each computing block could be easily visualized. It can be observed that

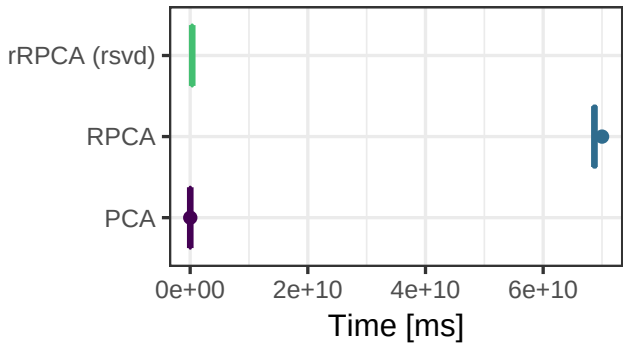


Figure A.1: Benchmark of PCA and RPCA algorithms.

RPCA with the exact ALM was the slowest, by a factor of more than 10^2 in comparison to the second slowest approach. It should also be noted that the times shown in Fig. A.1 for the exact RPCA were recorded from a custom compiled *R* package, linking against the Intel®MKL which provided a multi-threaded SVD implementation, whereas the rSVD was single threaded. Therefore, under the most popular scenario of vanilla *R* usage, without any custom optimizations, RPCA would have been up to $4 * 10^2$ slower that rRPCA, since a quad-core CPU was used in this case. With such a significant difference in computation time, the exact RPCA was deemed not viable.

In addition, classical PCA was included to demonstrate the impact of low-rank matrix recovery in rRPCA. This is shown in Fig. A.2.

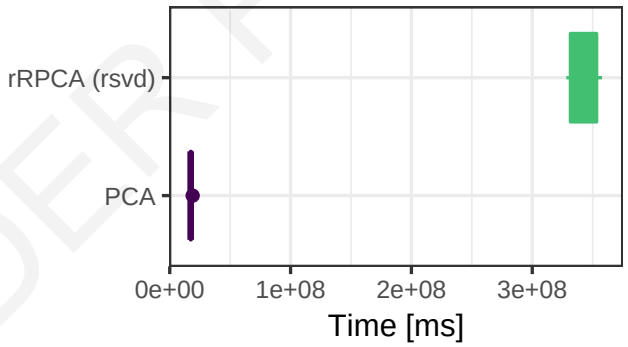


Figure A.2: Impact of low-rank matrix recovery with rRPCA.

In conclusion, from Fig. A.1 and Fig. A.2, it can be seen that the choice of the underlying algorithm providing RPCA had the largest influence on the overall performance of the data pipeline. Choosing the rRPCA was therefore the most logical step in reducing the computational cost of the unsupervised methodology.

Appendix B

Open-Source Contributions

B.1 Photovoltaic Performance Analysis in *R*

Photovoltaic Performance Analysis in *R* [41] (*pvpaR* for short) is a proof-of-concept web application that was developed for on-demand visualization and supervision of PV systems. It was released under the GNU General Public License (GPL), as a derivative of the work presented in this thesis. The application is a user-friendly online dashboard developed in *R* with features such as visualization of PV system performance, detection and classification of outliers and qualification of raw measurement data. The platform has been presented at the IEEE Photovoltaic Specialists Conference [41] and is freely available to download at `https://github.com/alexisph/pvpaR`.

Bundled with the application is a sample of data from the recorded operation of an actual PV plant at the testing site. The application is designed in such a way that it can be used with any data feed for rapid analysis and prototyping, and the data sample helps to showcase the application's functionality.

Current functionality includes visualization of measurement data and points of sub-optimal performance, irradiance modelling and transposition to the POA [216], modelling of PV system performance and quantification of differences between expected and actual performance. In addition, the application can detect and classify outliers. Loss parameters are estimated on the data given and are subsequently used to estimate acceptable performance thresholds. For points beyond the thresholds, a rudimentary classification and localization of the root cause procedure was implemented, loosely based on published research [76].

Future work will incorporate all methodologies developed and described in this dissertation. The envisaged architecture can be seen in Fig. B.1.

B.2 Optimization and packaging of *R* for Linux

All the results produced in the context of this dissertation were computed/estimated in *R*, an open-source software environment for statistical computing [181]. Similar to Python,

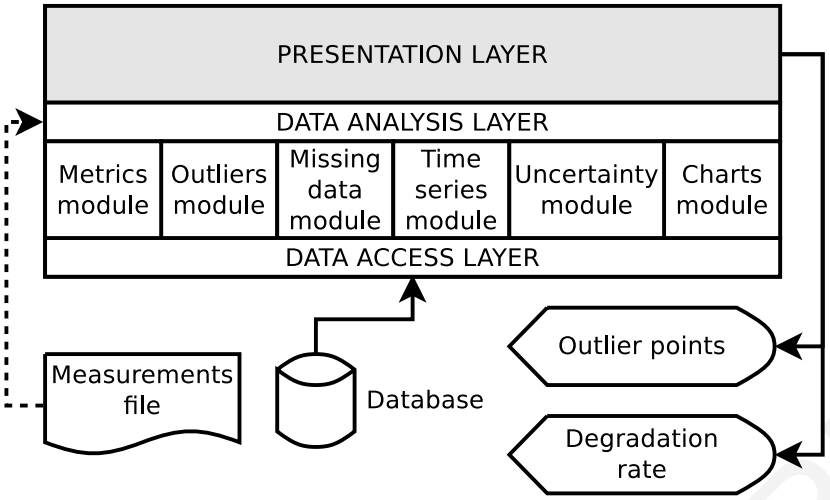


Figure B.1: Envisaged architecture of pvpaR.

Matlab, Julia and others, *R* is a statistical computing environment which includes libraries for data manipulation, calculation and graphical display. What sets it apart is its intuitive syntax and its vast library of open-source state-of-the-art statistical methods. *R* is very popular in both academia and industry.

Since statistical programming in *R* was a major part of this work, the base application was optimized as detailed below. In summary, the optimization procedure required custom compilation options and linking with high-performance, multithreaded mathematical libraries.

The configuration was hosted online in the Arch User Repository (AUR), which is a community-driven software repository for Arch Linux users. It contains package descriptions (`PKGBUILD`s) that allow anyone to install software via the package manager, without having to run `configure` and `make` scripts and having to manually keep track of installed files in the filesystem. The AUR was created to organize and share new packages from the open-source community and to help expedite popular packages' inclusion into the official repositories.

An Arch Linux user could fetch and install the software as customized for this work by downloading the `r-mkl` package and invoking `$ makepkg -si` at the terminal. When this is ran, the source code will automatically be downloaded, unpacked, compiled using the specific options and libraries as per the `PKGBUILD` and then installed. The package had already gained some popularity worldwide, as can be seen on the AUR page at https://aur.archlinux.org/packages/r-mkl/.

The package was compiled with the following flags and linked to the following libraries:

- `-O3`: enables aggressive optimization, including optimizations that incur a space-time tradeoff in favor of time, such as loop unrolling and automatic function inlining.
- `-xHost`: tells the compiler to generate instructions for the highest instruction set available on the compilation host processor, e.g. SSE or AVX.
- `-m64`: tells the compiler to generate code for Intel®64 architecture.

- `-qopenmp`: enables the parallelizer to generate multi-threaded code based on OpenMP directives.
- `-ipo`: enables interprocedural optimization between files.
- `-fp-model strict`: disables optimizations that are not value-safe on floating-point data and maintains ANSI/IEEE standards [217] compliance.
- `-fp-model source`: rounds intermediate results to source-defined precision.
- `-lpthread`: enables threading support.
- `-lm`: links the math library.
- `-lsvml`: links the Intel Short Vector Math Library.

Aside from the custom build of *R* described in the previous paragraph, the author also maintains an AUR package for Microsoft's enhanced, multi-platform distribution of *R*, called Microsoft R Open https://aur.archlinux.org/packages/microsoft-r-open/. This package also links against the Intel®MKL which it bundles alongside the binaries but does not enable some of the optimization options of the `r-mkl` package. Nevertheless, this distribution also provides much better performance than vanilla *R* and is currently the most straightforward way to run a multi-threaded analysis on Windows, Linux and Mac OS X. It can also prove useful when the Intel®MKL libraries cannot be licensed for free, or when the user does not want to install the Intel®development libraries package on their local machine for compiling `r-mkl`.