

Investigating solar farm siting criteria: A data-driven empirical method with case studies in Karnataka and Rajasthan

Student number: 100248602

Supervisor: Dr Amii Harwood

© This copy of the dissertation has been supplied on condition that anyone who consults it is understood to recognise that its copyright rests with the author and that use of any information derived there from must be in accordance with current UK Copyright law. In addition, any quotation or extract must include full attribution

Abstract

Increased Solar Energy Generation (SEG) will help to mitigate climate change. For optimal energy yield and economic performance SEG must be appropriately sited. This study analyses the location and capacity of 695 SEG plants in Karnataka and Rajasthan, two Indian states with extensive installed SEG, to achieve four goals:

- 1) Infer apparent criteria used for site selection and their relative importance
- 2) Find underlying, generalisable patterns in criterion importance
- 3) Generate maps of SEG suitability and make siting recommendations
- 4) Develop and test an empirical, data-driven method to achieve these goals.

Plant data is used to generate maps of SEG density, which are then modelled in terms of climate, terrain, social and infrastructure variables with multilinear regressions. Regressions are refined by testing for consistency, plausibility and effect size. These regression equations show the apparent relationships between the independent variables and SEG density, which are taken to reflect siting criteria, and used to model suitability.

The majority of criteria and weights determined were compatible with previous, expert-informed studies, suggesting that the method is viable and that plant siting is consistent with expert recommendations. Many variables appear to be more important where less favourable, possibly indicating a broader principle. For example, differences in levels of solar irradiance appear to have more impact on siting choices in areas that receive less sun.

Jaisalmer district in Rajasthan was found to be highly suitable and underexploited. Few such areas were found in Karnataka, suggesting more coherent siting there in terms of the variables assessed.

Key words:

Solar energy, Siting, Regression analysis, Determinants, India, Geographical information systems

Contents

1	Introduction.....	4
2	Literature Review	7
2.1	Determinant studies	7
2.2	Suitability studies.....	8
3	Method.....	11
3.1	Data used	11
3.1.1	SEG data	11
3.1.2	Climate	12
3.1.3	Terrain	13
3.1.4	Infrastructure	17
3.1.5	Social	17
3.2	Data transformations	18
3.3	Hypotheses	22
3.4	Regression testing	22
4	Results	27
4.1	Rajasthan	27
4.1.1	Backwards phase.....	27
4.1.2	Forwards phase	27
4.1.3	Mapping and Evaluation	28
4.2	Karnataka.....	31
4.2.1	Backwards phase.....	31
4.2.2	Forwards phase	33
4.2.3	Moving Window Regression.....	33
4.2.4	Mapping and evaluation	37
5	Discussion.....	41
5.1	Weight analysis.....	41
5.1.1	Climate	41
5.1.2	Terrain	47
5.1.3	Infrastructure	47
5.1.4	Social	48
5.2	Method Evaluation	49
6	Conclusions.....	51
7	Acknowledgements.....	52
8	References.....	53
9	Appendices	63

1 Introduction

Industrial era human activity has caused an increase in global temperatures of approximately 1°C (Allen *et al.* 2018). If warming continued at its pre-COVID rate we could expect to reach 2 °C of warming by 2070, with serious impacts on agriculture, sea level rise, biodiversity, and flood risk (Allen *et al.* 2018). CO₂ emissions are the key driver of this warming (Forster *et al.*, 2007: 137-152; Jain *et al.*, 2000). As around 85 % of CO₂ emissions come from burning fossil fuels (Friedlingstein, 2020) it is crucial that we reduce fossil fuel consumption even as energy demand rises by 4-9 % by 2030 (International Energy Agency (IEA), 2020a: 18). The potential of solar energy generation (SEG) for climate change mitigation has long been acknowledged (e.g. Arvizu *et al.* 2011); now, with mature technologies and falling prices, global installed capacity is expected to grow rapidly (IEA, 2020a: 18).

SEG can increase energy equity (Balta-Ozkan *et al.*, 2021) and access (Aly *et al.* 2017). Lack of access to electricity is associated with lower earnings, reduced study hours, and poorer medical facilities (Bensch *et al.* 2011; Aly *et al.* 2017). This impacts UN sustainable development goal 7: “access to affordable, reliable, sustainable and modern energy for all” (UN General Assembly, 2015).

These points are of particular relevance in India. National energy consumption rose 50 % from 2007-2017, and coal’s contribution has increased steadily since the 1970s (IEA, 2020b: 21, 229), yet up to 100 million Indians lack access to electricity (IEA, 2020b: 157)¹. At the same time, India has ambitious SEG policies: The 2008 National Action Plan on Climate afforded solar a central role (Jawaharlal Nehru National Solar Mission (JNNSM), 2009), and states are obligated to procure increasing renewable energy each year (Yenneti, 2016). The JNNSM has exceeded a series of multi-annual targets for national SEG capacity. It has fostered policies to promote SEG, including off-grid, rooftop (RTPV), “ultra-mega solar parks” (UMSPs), and solar irrigation (Yenneti, 2016; Adhikari, 2020; Shrimali and Rohra, 2012) (glossary of terms in [Table 1](#) at end of Introduction). Installed SEG capacity rose from <1 GW in 2012 (International Institute for Sustainable Development, 2017) to 32 GW of PV alone in 2019 (IEA, 2020b: 107). The Central Electricity Authority (CEA) predicts 280 GW of SEG by 2030 (CEA, 2020: 16).

For India to derive maximum benefit, additional capacity must be optimally located. This is the role of siting studies: to determine areas most suitable for SEG, in terms of energy output, economic viability, and other factors (e.g. Sánchez-Lozano *et al.* 2013, Kengpol *et al.* 2013, Sindhu *et al.* 2017). Improved efficiency reduces per-unit negative impacts, including pollution associated with extraction and disposal of panel materials (Hertwich *et al.*, 2015; Corcelli *et al.*, 2017).

¹ Upper estimate. Other sources suggest figures from six million (IEA, 2020c) to 4.8 % of the population, ≈ 65 million (IEA and World Bank, 2018)

The utility of a siting study depends on appropriate choice of exclusion and evaluation criteria, and appropriate weighting of evaluation criteria. Exclusion criteria determine where sites cannot be constructed, due for example to inaccessibility (e.g. Merrouni *et al.* 2018) or excessive slope (e.g. Uyan, 2017). Evaluation criteria are used to calculate suitability. They typically include measures of irradiance and proximity to transmission infrastructure and population centres, but otherwise vary widely.

The literature review, below, will show that all suitability studies reviewed for this report depend on expert or authorial opinion for determination of weights. In the terminology of location theory, these are prescriptive, not descriptive studies: they make recommendations, rather than analysing the current situation (Chan, 2011: 83). A separate body of studies analyses site locations to find determinants, but typically focus on residential RTPV adoption patterns. None of the determinant studies reviewed analyse suitability. There appears to be a paucity of studies combining descriptive and prescriptive elements: analysing current SEG to infer criteria, and using that information to create suitability indices and make recommendations.

Such combined analysis is the goal of this study. While the studies reviewed below typically draw on the knowledge of 2-10 experts, of whom in several cases none are industry professionals, the plants in this study's dataset embody decisions taken by thousands of local SEG professionals. Most locations are chosen from multiple options. Even when this is not the case, as with SEG installed by commercial or industrial consumers for on-site electricity ("distributed" SEG), locations are assessed for viability. In all cases a decision is made about capacity to install. One would therefore expect a positive correlation between suitability and capacity.

This may be obscured if siting choices do not reflect broadly applicable criteria, or reflect criteria other than those analysed. Choices may be influenced by localised effects, such as local policy or business relationships. The construction of one site may favour the development of more. If a region's workforce gains relevant skills during the construction of a UMSP, other site planners may be more likely to build there, independent of the area's suitability in other regards. Such "agglomeration economies" (Arthur, 1994: 35; Seppaelae, 1997) could cause misleading correlations.

This study's central question is whether regression analysis of the location, capacity and characteristics of existing SEG can nevertheless provide insight into siting criteria, and allow the construction of empirically grounded, regionally customised suitability indices, in contrast to the more common grounding in expert opinion. To provide a rich dataset, the two Indian states with most installed SEG were selected: Karnataka and Rajasthan ([Figure 1](#)). Karnataka has the greatest current capacity, and thus provides the richest dataset, but Rajasthan has much more planned, so

understanding criteria there may be more important in future. Rajasthan has the greatest availability of land that is highly suitable for SEG of the 36 Indian states, while Karnataka ranks 4th, according to (Saraswat *et al.*, 2021). Karnataka's SEG is more evenly distributed than Rajasthan's, making it more mathematically challenging to extract information from the Rajasthan data.

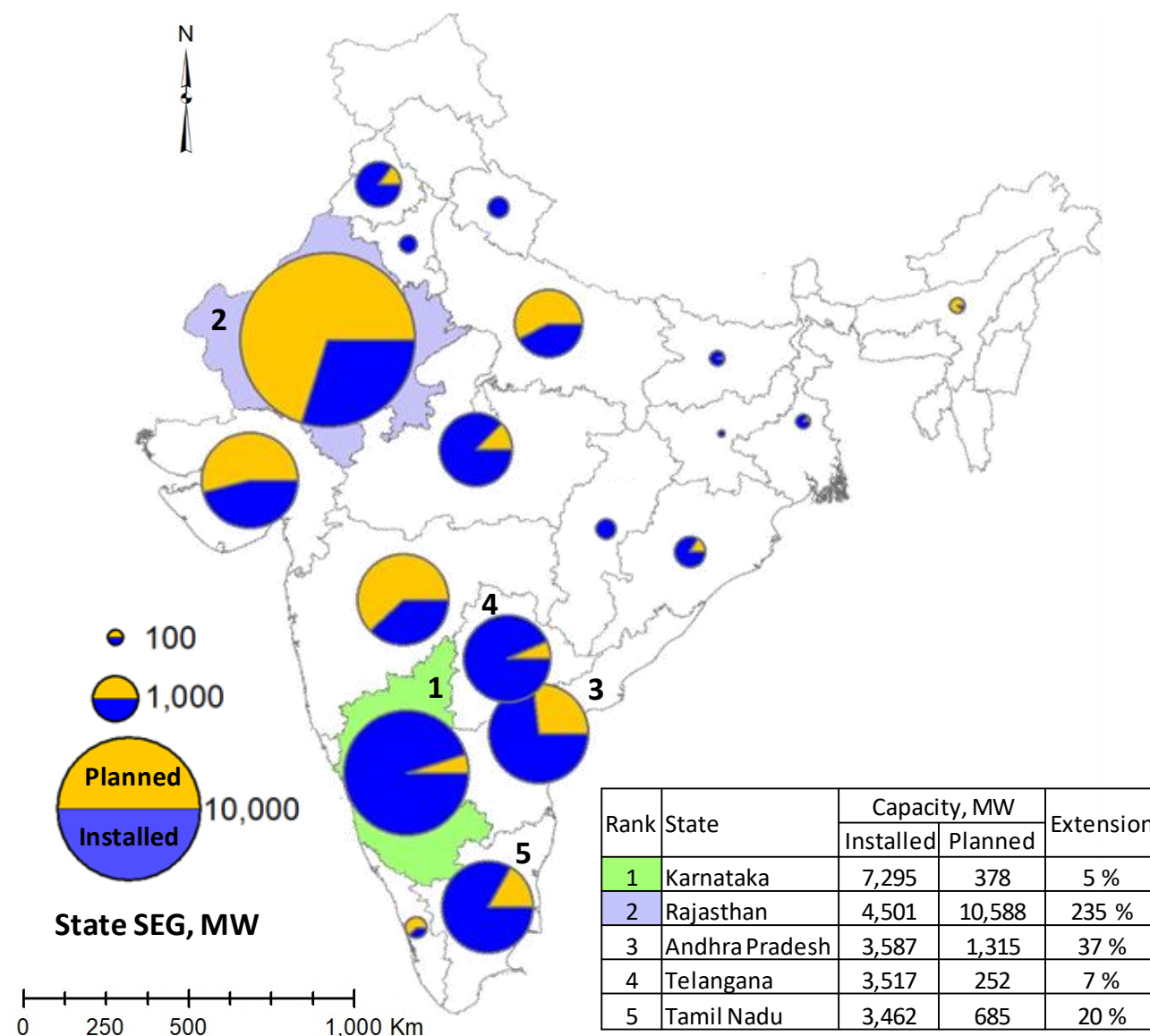


Figure 1: Planned and installed SEG capacity by state. Inset: top five states by installed SEG. "Extension": planned capacity as % of installed. Rajasthan's planned SEG outstrips any other state. All mapping and spatial processing used ArcGIS 10.6.1 (Environmental Systems Research Institute (ESRI), 2018).

Table 1: Glossary. † indicates economic terms that will be more narrowly defined as variables

AHP	Analytical Hierarchy Process	LULC	Land Use, Land Cover
AODb	Aerosol Optical Depth, blue (dust proxy)	MCDA	Multi-Criterion Decision Analysis
β	Regression coefficient	MLR	Multiple Linear Regression
β^*	Standardised regression coefficient	MWR	Moving Window Regression
GIS	Geographical Information Systems	PCI	Per Capita Income †
GPB	Grampanchayat Budget †	RTPV	Roof-top PV
GWR	Geographically Weighted Regression	SEG	Solar Energy Generation
HSTR	Homogeneous Smooth Transition Regression	TDP	Taluk-level Domestic Product †
ISC	Indian Subcontinent	UMSP	Ultra-Mega Solar Park

2 Literature Review

2.1 Determinant studies

A search was conducted for studies of SEG determinants using the search term “photovoltaic determinants”. All searches were conducted on a combined database with access to the Web of Science, ScienceDirect, Scopus, IEEE Xplore and EThOS databases. Of the first hundred results, fifteen were empirical studies of SEG location (Appendix [9.1](#))². None of these created a suitability index. 12 considered only domestic RTPV. Of the remainder, one also considered community generation (Schunder *et al.*, 2020), and two investigated distributed generation (Peter *et al.*, 2006; Lin *et al.*, 2017). Four focussed on part of the Indian subcontinent (ISC), but only one on India (Peter *et al.*, 2006).

All except (Lin *et al.*, 2017) are adoption studies, analysing the characteristics of the adopter, though some also consider environmental factors. An emphasis on social criteria is common, including education (Jan *et al.*, 2020; Schunder *et al.* 2020; Balta-Ozkan *et al.* 2015), sociodemographic characteristics (Dharshing *et al.*, 2017; Copiello and Grillenzoni, 2017; Lee *et al.* 2019), and householder political preference (Briguglio and Formosa, 2017). Copiello and Grillenzoni (2017) consider climate and social factors and find that RTPV distribution is only weakly related to climate.

10 studies performed regression analysis, of which 7 featured multiple linear regression (MLR). A MLR explaining dependent variable y in terms of K independent variables $x_1...x_K$ has the form

$$y_i = \beta_0 + \sum_{k=1}^K \beta_k x_{ki} + \epsilon_i$$

Equation 1

At the i th datapoint, y is equal to an invariant constant β_0 , plus the sum of the x terms $x_1...x_K$ at i each multiplied by its respective regression coefficient $\beta_1... \beta_K$, plus error term ϵ (Fotheringham *et al.* 2002: 52; Gujarati, 2019). The regression coefficients that minimise $\sum \epsilon^2$ best approximate the relationship between the dependent and independent variables (Pek *et al.*, 2018).

Balta-Ozkan *et al.*, (2021) use MLR to identify determinants, and geographically weighted regression (GWR) to analyse them. GWR recognises that relationships may vary spatially (Fotheringham *et al.* 2002: 9) using [Equation 2](#):

² For ease of document navigation, all Appendices include hyperlinks to return to the point of origin within the report.

$$y_i = \beta_0(u_i, v_i) + \sum_{k=1}^K \beta_k(u_i, v_i)x_{ki} + \epsilon_i$$

Equation 2 (Fotheringham *et al.* 2002: 52)

u_i, v_i are the coordinates of the i th point. For each coefficient, the continuous function $\beta_k(u, v)$ describes its spatial variation.

A search for “geographically weighted regression solar” yielded no further relevant results.

2.2 Suitability studies

Searches were conducted for “(solar OR photovoltaics) location analysis” and “(solar OR photovoltaics) siting criteria multilinear regression”. The first hundred results for each were filtered further. 49 were relevant. Of these, 29 were excluded because they focussed on single issues. Snowball sampling of the remaining 20, together with subject-specific searches, yielded 22 further studies.

Of the 42 papers, 79 % used geographical information systems (GIS) (Appendix 9.2). 76 % used Multi-Criteria Decision Analysis (MCDA)³. 69 % used both. MCDA encompasses various techniques that facilitate decision-making in situations with conflicting objectives (Mardani *et al.*, 2015a). It is often used in the field of sustainable energy (Mardani *et al.*, 2015b). Of studies that specified a form of MCDA, 88 % (28 papers) used some form of Analytical Hierarchy Process (AHP).

AHP starts with criteria that are relevant to the decision and, within those, subcriteria. Pairwise comparisons between criteria, and between subcriteria within each criterion, are used to generate weights (Saaty, 1994). Weights indicate the importance of each subcriterion and can be used to make choices, create a suitability index, or, in conjunction with GIS, generate a suitability map (e.g. Merrouni *et al.*, 2018). Although AHP uses a mathematical process to generate weights, the input is comparisons made by individuals, introducing subjectivity (Tavana *et al.*, 2017; Waewsak *et al.*, 2020). This may yield flawed results if decision-makers exhibit bias (Kengpol *et al.* 2013; Uyan, 2017). It is recommended that comparisons be performed by experts (Saaty, 1994).

Of 42 siting studies analysed, 11 stated that comparisons were performed by external experts, 1 by “energy related groups” (Giamalaki *et al.*, 2019), and 5 that discussion with experts informed comparisons made by the authors. 25 did not mention external consultation. In 8 there was a defined role for external SEG professionals. Some authors validated findings against other research on their study area (e.g. Jun *et al.*, 2014) or against SEG location (e.g. Palmer *et al.*, 2020), or created multiple scenario-specific indices (e.g. Azevêdo *et al.*, 2017; Ruiz *et al.*, 2020).

³ Statistics given here only consider a study to have used MCDA if the study method mentions such use.

Frequency of inclusion of selected criteria is presented in [Table 2](#). All studies gave criteria for photovoltaics (PV) except one which focussed on Concentrating Solar Power (CSP) (Azevêdo *et al.*, 2017).

Suitability studies inform decisions taken by investors (Akkas *et al.*, 2017; Ruiz *et al.*, 2020) and government (Saraswat *et al.*, 2021). Many criteria are therefore discussed in economic terms: proximity to roads reduces transport costs (Ghose *et al.*, 2020), proximity to power lines reduces connection costs (Doljak *et al.*, 2017), low slope reduces construction costs (Uyan, 2017), panel soiling through dust deposition raises cleaning costs (Charabi and Gastli, 2012). Nonetheless, only five studies include direct economic factors, focussing on costs (land lease fees, operation and maintenance, etc.) (Appendix 9.3). Although government policy is relevant (Lin *et al.*, 2017; Doljak *et al.*, 2017), only two studies include it in their suitability indices (Samanlioglu and Ayağ, 2017; Sindhu *et al.*, 2017).

Table 2: Frequency of criterion inclusion in studies reviewed
Criteria assigned 0 weight are not included. “+” indicates that higher values or greater proximity are considered beneficial to SEG, “-” detrimental, and “±” mixed effects or no consensus. “\” indicates category data or variable impact. A criterion used in one study for both exclusion and evaluation is counted as evaluation. Visual rounding may lead to inconsistent totals. “Population” refers to population density or proximity to populated areas. Full data: Appendix 9.2)

Category	Criterion	% Papers including			Influence	Reason
		Exclusion	Evaluation	Total		
Climate	Irradiance	2	90	93	+	Increases energy generation
	Temperature	2	29	31	-	Reduces efficiency
	Panel soiling (dust)	7	7	14	-	Reduces incident sunlight (ISL)
	Humidity	0	10	10	-	Reduces ISL
	Wind	2	5	7	±	Cooling effect; hazard
Terrain	LULC*	40	36	76	\	Affects legality and ease of construction
	Water	40	21	62	±	Used for cleaning; hazard
	Slope	10	50	60	-	Affects ease of construction
	Aspect	5	36	40	\	Orientation towards equator increases ISL
	Elevation	5	17	21	±	Cooling effect; construction difficulties
	Available Area	10	5	14	+	Large continuous areas facilitate utility-scale construction
	Hazards	5	5	10	-	Long-term threat (e.g. seismic faults)
Infra-structure	Roads	2	69	71	+	Proximity facilitates access
	Powerlines	7	57	64	+	Proximity facilitates transmission
	Substations	0	12	12	+	Proximity facilitates transmission
Social	Population	7	45	52	±	Various impacts; see text
	Economic factors	0	12	12	\	Various impacts; see text
	Policy	0	5	5	\	Various impacts; see text
	Demand for electricity	0	0	0	+	Proximity to demand centres reduces transmission costs

* Land Use / Land Cover

While no studies include direct measures of electricity consumption in their indices, several include proxies. Aly *et al.*, (2017) use proximity to mines due to the extensive mining activity in their study area, (Jun *et al.*, 2014) use GDP, (Kengpol *et al.*, 2013) use distance to “locations which

demand power”, and others discuss distance to population centres in the context of electricity demand or transmission losses (e.g. Al Garni *et al.*, 2017a; Colak *et al.* 2020; Merrouni, 2018). Others recommend maximising this distance. Reasons include nuisance (Saraswat *et al.*, 2021), NIMBYism (Sindhu *et al.*, 2017), urban air pollution causing panel soiling (Tercan *et al.*, 2020), and future urban expansion (Azevêdo *et al.*, 2017). Of 20 papers that provide weights for a population criterion, 12 consider it beneficial and 8 detrimental ([Figure 21](#), Discussion). Of those that evaluate population negatively and state distance bands used, the median value for the bottom of the top band was 10 km, with the third quartile value being 43 km. This bottom-of-top figure is the value above which additional distance does not alter modelled suitability. For slope, median and third quartile bottom-of-top values were 5.1° and 13.5° (22 studies, Appendix [9.2](#)).

3 Method

This section will first introduce the data used. All frequently assessed criteria ([Table 2](#)) will be discussed, with reasons for use or rejection. Hypotheses will be outlined and regression methods presented.

3.1 Data used

3.1.1 SEG data

SEG capacity and location data were obtained for Karnataka (to 03/2019) and Rajasthan (to 05/2020) from government and industry sources respectively (sources for all data used in report: Appendix [9.4](#)). Location was in the form District, Taluk, Village. Taluks are administrative divisions also called sub-districts (e.g. Directorate Of Census Operations Andhra Pradesh 2011: 26). The administrative division “village” does not connote rurality.

District-taluk-village SEG locations were cross-referenced with GIS data, using Fuzzy Lookup v1.3.3.0 (Microsoft Research, 2020) to return the closest match. This was necessary due to variant transliterations. Location matches with <90 % similarity were excluded. Matches <95 % were checked. 356 plants totalling 6527 MW were located to village level in 272 villages in Karnataka, and 339 plants totalling 4922 MW in 88 villages in Rajasthan, representing 89 % and 109 % respectively of installed capacity as of December 2019 (Bridge to India, 2020a) ⁴. If SEG villages are modelled as squares of consistent size, this equates to locating each plant in a cell on a grid of resolution equal to the square root of the mean village area. Mean root area for SEG villages is 3.8 km (standard deviation $\sigma = 1.8$ km) for Karnataka and 6.5 km ($\sigma = 4.1$ km) for Rajasthan. This indicates the approximate linear scale of precision.

For each village, total capacity and capacity per unit area are influenced by the administrative boundaries. If boundary positions changed, patterns in the data would change. This is a form of the Modifiable Unit Area Problem (MAUP), one response to which is to interpolate data spatially (Wong, 2009). SEG data was therefore imported as capacity values in MW at village centroids and smoothed using the kernel density function. This fits a smooth surface over point data. For any datapoint, density is greatest at the point and diminishes to zero at a specified radius. Mean density is equal to point value divided by area (Appendix [9.5](#)). SEG density units are $\text{MW}\cdot\text{km}^{-2}$. As the majority of criteria being considered vary continuously it is assumed that, if SEG has been installed at a given location, nearby locations will be somewhat suitable. Implied suitability is proportional to installed capacity, and confidence diminishes with distance. The SEG density layer mimics this, interpolating modelled suitability as a function of installed capacity and

⁴ The higher percentage for Rajasthan is due to more recent data.

distance. Kernel density radius was set to 15 km as a balance between setting it too high and including as suitable land entirely removed from known sites or setting it too low and providing an insufficient dataset for MLR. All data layers were reprojected in the Kalianpur 1975 Universal Transverse Mercator Zone 43N projected coordinate system. This is the recommended system for the study area (ESRI, 2012: 25).

3.1.2 Climate

Irradiance is considered as an evaluation criterion in 90 % of studies assessed, and was therefore evaluated⁵. Measures of irradiance used include Direct Normal Irradiance (DNI) (Suh and Brownson, 2016; Tahri *et al.*, 2015) and Global Horizontal Irradiance (GHI) (Tercan *et al.*, 2020; Mensour *et al.* 2019). DNI measures solar radiation received directly, while GHI also includes Diffuse Horizontal Irradiance (DHI). DHI is the component of shortwave energy which arrives at the collector after being scattered by the air. DNI or GHI data may be preferred for planning a particular plant because CSP harnesses only direct irradiance, while PV uses both direct and diffuse (Giamalaki *et al.*, 2019). PV predominates in India as a whole, though CSP represents around 50 % of Rajasthan's capacity (Kumar *et al.*, 2017). However, the high penetration of CSP in Rajasthan reflects the high proportion of solar energy arriving as DNI. It was concluded that GHI would not problematically overstate the solar resource in Rajasthan, but would understate it in Karnataka, so GHI was chosen. As solar collectors are not usually horizontal it is necessary to correct for inclination (Brownson, 2014: 211). Global Tilted Irradiance (GTI) modifies GHI for tilt angle (Jessen *et al.*, 2018). GTI data used here assume optimal tilt angle (Energy Sector Management Assistance Program, 2019). (All evaluation layers: [Figure 6](#), [Figure 7](#)).

Temperature is evaluated in 29 % of studies. An increase of 1 °C reduces PV efficiency by ~0.45 % points (Skoplaki and Palyvos, 2009), which may be material if baseline efficiency is low. Temperature was therefore evaluated. Panel soiling is only considered in 14 % of studies, but Rajasthan's desert regions make soiling from dust deposition a concern (Ginley *et al.*, 2020). Soiling can reduce energy reaching panels by 10 % per month (Laarabi *et al.*, 2020); reductions of 70 % have been recorded (Albadwawi *et al.*, 2019). Aerosol Optical Depth in the blue wavelength band (AODb) is used as a proxy for atmospheric dust concentrations for evaluation of dust deposition risk on solar panels (Charabi and Gastli, 2012; Charabi and Gastli, 2013). AODb was therefore evaluated.

Daily AODb satellite data for 2013-2019 were obtained and collated by season using Python 3.8 (Van Rossum and Drake, 1995) (Appendix [9.6](#)). Monsoon period data were disregarded as dust would not accumulate on panels. Datasets were large ($n_{\text{points}} > 10^6$ per state) but coverage

⁵ "Evaluated" is used to mean included as an evaluation criterion, not merely for exclusion.

variable, so 10 km smoothing was applied with mean focal statistics. Values were averaged with equal weight given to each season and year to prevent sampling bias.

Only 4 out of 42 studies assess humidity, and only one of these focusses on the ISC. Three studies consider wind, and none on the ISC. Therefore neither wind nor humidity was considered.

3.1.3 Terrain

LULC is an evaluation criterion in 36 % of studies and an exclusion criterion in 40 %. To evaluate LULC, suitability values must be assigned to each LULC category, which was considered undesirable due to the expert knowledge required. To empirically inform LULC exclusion criterion selection, representation was compared in each state as a whole and in SEG villages ([Table 3](#)).

Village proportions were capacity-weighted using [Equation 3](#).

$$\bar{P}_j = \frac{\sum_{i=1}^n \frac{A_{ij}}{A_i} C_i}{\sum C}$$

Equation 3

\bar{P}_j = capacity-weighted area proportion of LULC category j summed across SEG villages

n = number of SEG villages in the state

A_{ij} = area of category j in village i

C = capacity of SEG in village i

$\sum C$ = capacity in all SEG villages.

Conceptually, this is the proportion of state SEG that would be on land of each category if, in each SEG village, SEG were spread evenly. As this is not the case figures must be treated with caution.

Table 3: percentage of state and capacity-weighted SEG village land per category
“Over” and “Under” show the factor by which each category is over- or under-represented in SEG villages relative to the state as a whole. Over-representation factor = %_{SEG} / %_{state}; under-representation factor is the reciprocal of this. This is susceptible to skewing when sample size is small, e.g. with Wasteland and Plantation in Rajasthan. Deeper colours show larger values.

Karnataka		Proportion, %		Representation factor	
LULC category	State	SEG villages	Over	Under	
Barren	0.57	0.54		1.1	
Built-up	1.3	0.15		8.9	
Cropland	58	83	1.4		
Fallow	9.7	11	1.1		
Forest	15	1.1		13.6	
Grassland	0.00	0.00			
Plantation*	9.5	1.2		8.2	
Shrubland	3.1	1.3		2.4	
Wasteland	0.06	0.59	9.3		
Water bodies	2.9	1.5		2.0	

Rajasthan		Proportion, %		Representation factor	
LULC category	State	SEG villages	Over	Under	
Barren	10	40	3.8		
Built-up	0.64	0.33		2.0	
Cropland	35	13		2.6	
Fallow	30	27		1.1	
Forest	5.0	1.4		3.6	
Grassland	3.6	0.62		5.9	
Plantation*	0.05	0.00		889	
Shrubland	12	16	1.4		
Wasteland	1.5	0.03		53	
Water bodies	1.7	0.51		3.4	

* Plantation refers to land for tree crops (DAAC, 2016), which tend to embody higher capital.

Forest is frequently excluded or discouraged (e.g. Zoghi *et al.*, 2017; Giamalaki *et al.*, 2019; Uyan, 2017), and is underrepresented in SEG villages in both states. It was therefore excluded. Agricultural land is frequently excluded (e.g. Ruiz *et al.*, 2020; Mierzwiak and Calka, 2017) or negatively weighted (e.g. Sánchez-Lozano *et al.*, 2013; Chen *et al.*, 2014). In Karnataka cropland is slightly overrepresented, with 1055 MW of SEG is in villages >99.5 % cropland (n = 44). In Rajasthan cropland is somewhat underrepresented, with 15 MW in villages >99.5 % cropland (n = 2). This suggests that SEG developers in Rajasthan avoid cropland more than in Karnataka, but that this avoidance is not total. It may also weaken: Rajasthan has >10 GW of SEG planned, double current capacity, and recent state policy provides for the acquisition of agricultural land for SEG development beyond previous limits (Energy Department, Government of Rajasthan (EDGoR), 2019: 24). Cropland was therefore not excluded. Plantation land is heavily underrepresented in both states and was excluded.

Built-up land is generally excluded (e.g. Rediske *et al.*, 2020; Waewsak *et al.*, 2020; Merrouni *et al.*, 2018), and is underrepresented in both states. However, Karnataka and Rajasthan had 482 MW and 389 MW of non-residential RTPV respectively at the end of 2019 (Bridge to India, 2020b). In Karnataka this is more than double the total one year earlier, primarily in utility-scale projects (Karnataka Electricity Regulatory Commission (KERC), 2019a). This is far behind the 2021 target of 2,400 MW and the Government of Karnataka is taking regulatory steps to encourage faster growth (KERC, 2019b). Rajasthan has a target of 1,000 MW RTPV by 2025, with 300 MW in cities (EDGoR, 2019: 14, 16). Built-up land was therefore not excluded.

Legally protected wildlife and nature reserves were excluded, as were water bodies. Proximity to water bodies to facilitate cleaning was not evaluated because the underrepresentation of water bodies in SEG villages suggests this is not a priority.

To gain insight into local practice, mean slope values were taken for each SEG village and capacity summed in 0.5° bands ([Figure 2](#)). This is indicative only as it is not known where in the village SEG is located, but intra-village variation is low: at 500 m resolution the mean coefficient of variation, a measure of variation equal to standard deviation divided by mean, is 0.45 in SEG villages in Karnataka and 0.52 in Rajasthan. Slope values were generated in ArcMap from 30 m resolution elevation data. [Figure 2](#) shows a preference for shallower slopes in both states, but in Karnataka there is 128 MW of capacity across several sites at 5.5-8°. Since the states have similar overall slope profiles, and in recognition of Rajasthan's planned capacity increase, the upper limit was set at 10° and slope evaluated. State inclusion zones were created by excluding forest, plantation and water LULCs, protected land, and land with slope ≥10° ([Figure 3](#))

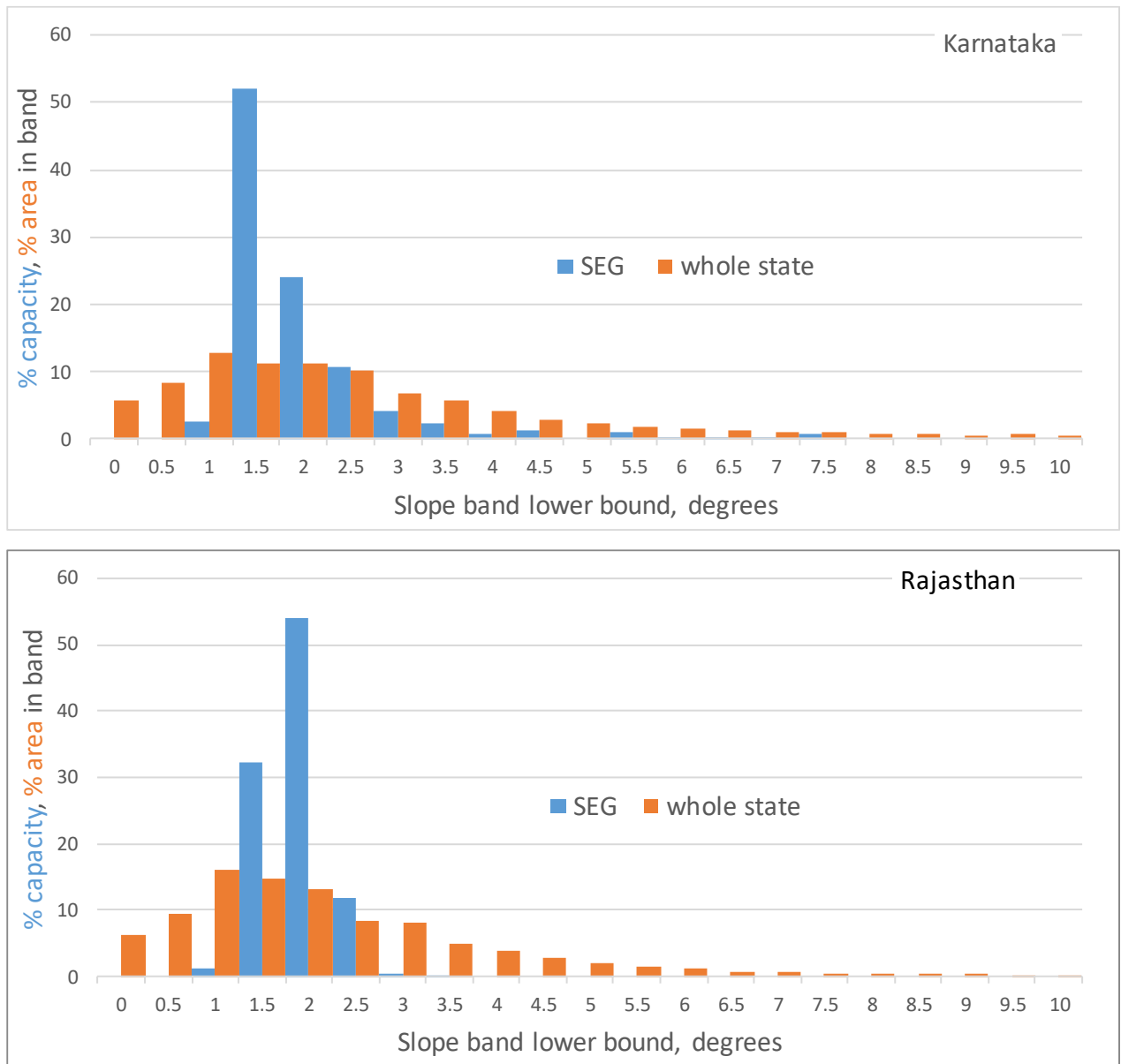


Figure 2: Slope frequency in SEG villages and state for Karnataka (top) and Rajasthan. Values > 10° not shown.

Aspect was not considered due to the difficulty of application at village-level resolution. Elevation is only considered in one Indian subcontinent study, where it is afforded low importance (Saraswat *et al.*, 2021), so was not analysed. The hazard most discussed in the literature is earthquakes (Colak *et al.*, 2020; Kengpol *et al.*, 2013; Tercan *et al.*, 2020), but these are not mentioned in any of the ISC studies reviewed. Hazards were therefore not considered. Available area is not considered in any ISC studies, and was therefore rejected.

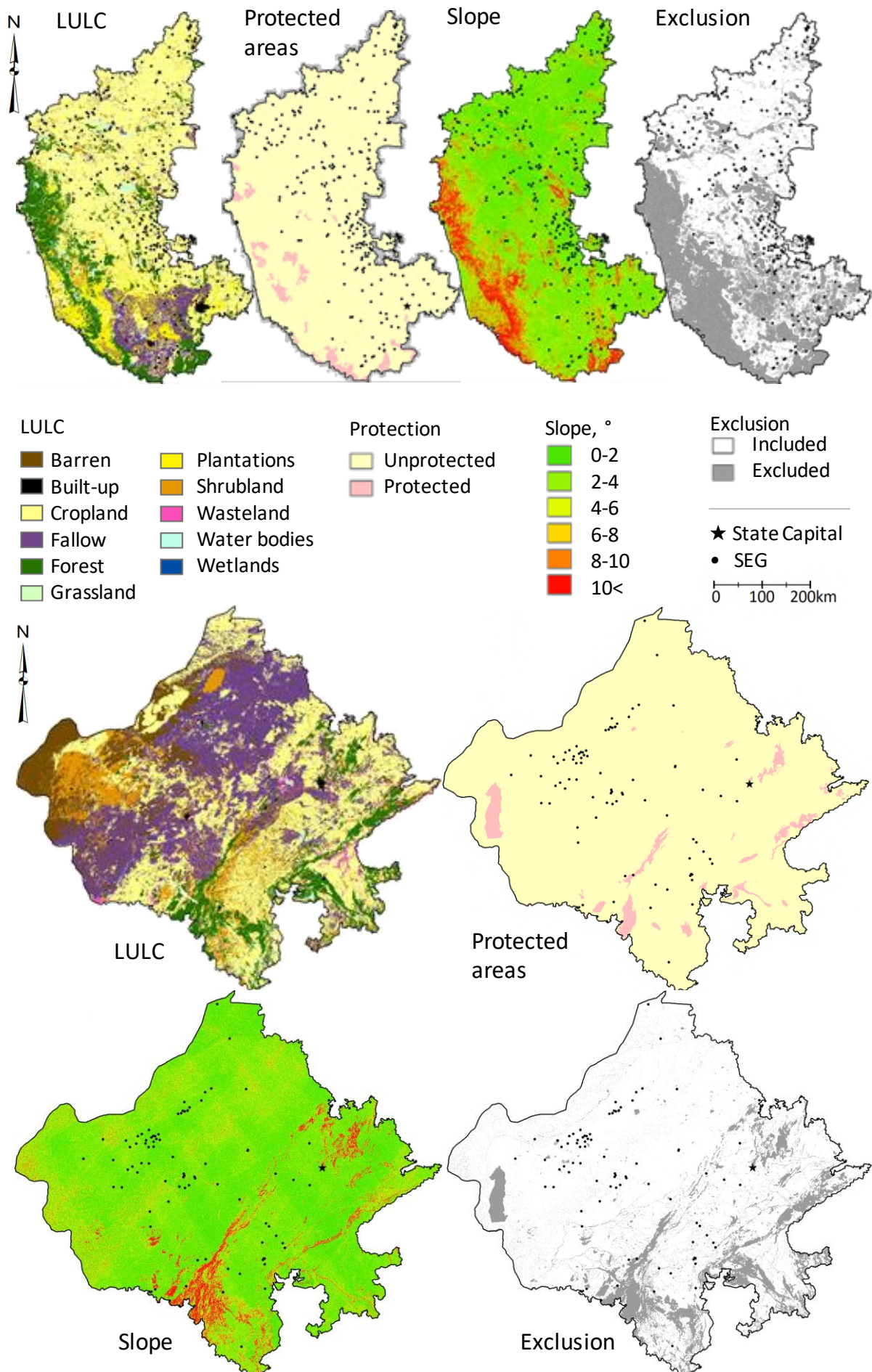


Figure 3: Exclusion criteria/zones for Karnataka (top) and Rajasthan. Key and scale apply to both. Note SEG on Karnataka cropland, and low slope in most SEG areas.

3.1.4 Infrastructure

Roads and powerlines are considered in most studies, including five out of six ISC studies (Appendix [9.2](#)), and were therefore evaluated. Distance from roads was calculated for all points. Regressions using different road data sources gave conflicting results in Karnataka, and were therefore rejected. For grid connection, both proximity and voltage of powerlines are important, so powerline kernel density was used as the Powerlines variable. Line density values are a function of $(\text{line length}) \cdot (\text{line value}) / \text{area}$, giving units of $\text{MV} \cdot \text{km} \cdot \text{km}^{-2}$. This can be thought of as a measure of the voltage-weighted length of powerlines running through an area. Most SEG villages lie well within 10 km of a powerline, so this was used as the limiting radius.

For large sites new powerlines may be constructed, such that present-day co-occurrence of powerlines and SEG does not represent the conditions when the site was chosen. This reduces quality of insight into original siting criteria. However, inspection of planning documents suggests that even for the largest sites this is generally a case of strengthening existing powerline corridors, rather than creating new ones, such that powerline density can still be considered the causal component in the case of powerline-SEG density correlation (e.g. Rajasthan Rajya Vidyut Prasaran Nigam Limited, 2012: 10).

Substations were not considered due to difficulties obtaining comprehensive data.

3.1.5 Social

Population density was calculated and averaged over various radii to provide smooth surfaces and take account of proximity effects. Annular averages were used as separate independent variables to test the effects of population in the immediate or more remote vicinity. SEG siting decisions were made in the context of past population levels, but suitability indices constructed in this report must be applicable in the future. To balance these issues, 2020 population data were used.

Land price is a key component of initial capital for utility-scale SEG (Wu *et al.*, 2014). It was hypothesised that land price would be anticorrelated with SEG density. As land price proxies, 2008-2009 Per Capita Income (PCI) and Taluk-level Domestic Product per km^2 (areal TDP) were mapped for Karnataka and 2018-2021 Gram Panchayat Budget per km^2 (areal GPB) for Rajasthan. Gram Panchayats are an administrative level between taluk and district. GPB was averaged at taluk level due to incomplete coverage. The age of the Karnataka data, the resolution of the Rajasthan data, and the different measures for the two states may limit utility.

Policy was not evaluated due to difficulties of quantification.

A web scraper harvested hourly real-time power consumption (demand) levels for 120 high-level substations from the Karnataka Electricity Supply Company monitoring website from 16/9/2020 to 30/10/2020 (code: Appendix [9.7](#)). A kernel density surface, radius 100 km, was used

to map power density accounting for proximity. For Rajasthan, 2018 consumption data were obtained at subdivision level ($n = 314$). Subdivisions were geocoded to generate coordinates (Appendix 9.8) and 100 km radius kernel density smoothing applied (Figure 4). Regressions were not run on this data due to incompleteness.

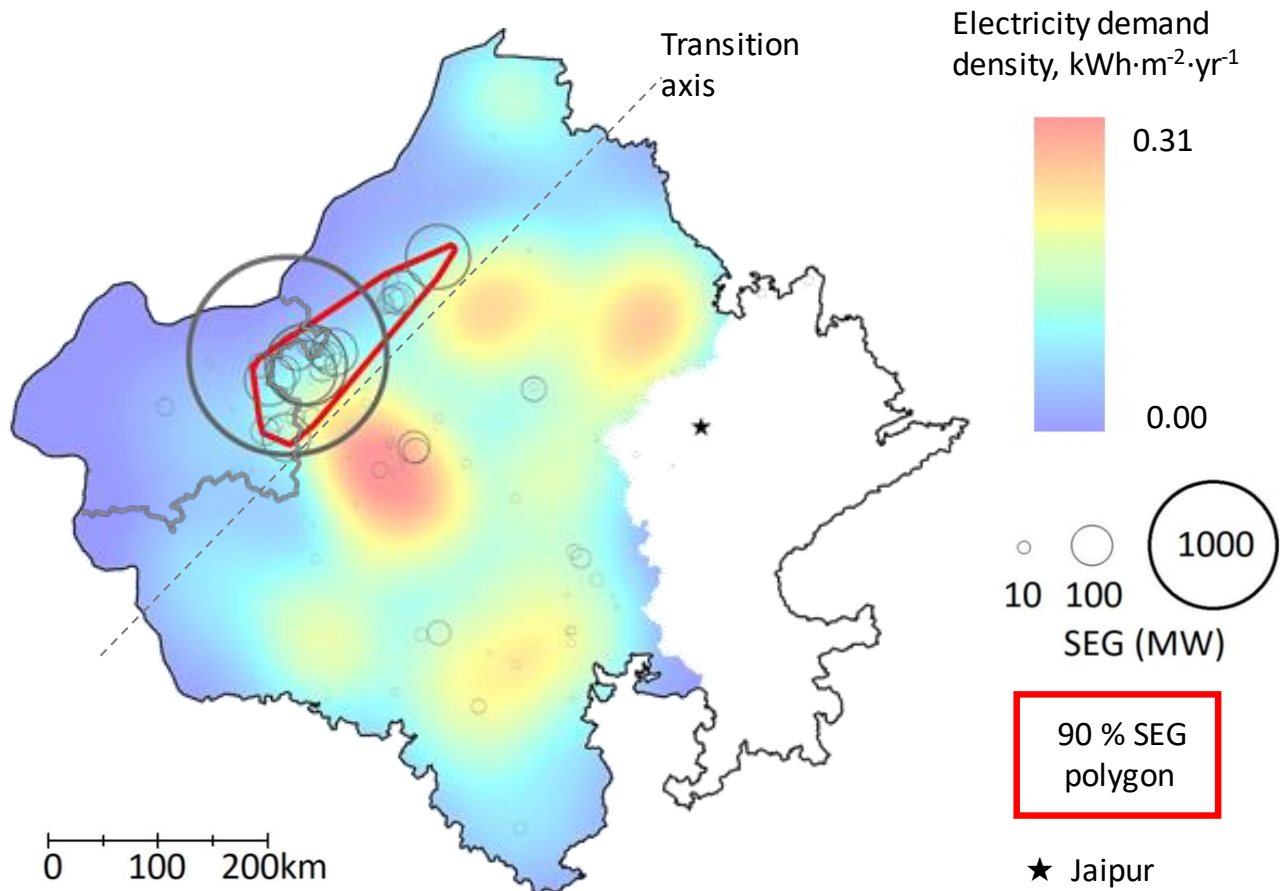


Figure 4: Electricity demand, Rajasthan. Note transition between high-demand south/east and low-demand west, including extremely low-demand Jaisalmer district (outlined grey). The red polygon contains 92 % of state SEG capacity, 3.5 % of state area. SEG villages represented by grey circles centred on village centroids. Circle area is proportional to total installed capacity in village.

3.2 Data transformations

Non-normally distributed input data can give misleading standardised regression coefficients (β^* , described below) (Greenland *et al.*, 1991). It was found that SEG density, population density and all land price proxies were severely non-normal according to criteria in (Kline, 2016: 76). These variables were base-10 log-transformed, and the resultant datasets conformed much more closely to normality criteria (Appendix 9.9). These independent variables were replaced with their log-transforms. As variable names, “Population” refers hereafter to $\log(\text{population density})$, “GPB” to $\log(\text{areal GPB})$, and “TDP” to $\log(\text{areal TDP})$. Population was initially averaged at 10 km, the median bottom-of-top-band value from the literature review.

Log transforming SEG density caused a reduction in available data because $\log(0)$ is undefined, excluding all land that is not within 15 km of an SEG village and therefore has SEG

density 0. In Karnataka the reduction was 44 %, so regressions were run on both original and log-transformed SEG density. These will be referred to as “non-log” and “log” regressions; all use MLR. In Rajasthan the reduction was 87 % ([Figure 5](#)), so non-log regressions were prioritised and, rather than repeat the backwards/forwards process for the log data, the combinations of variables found to be optimal in non-log regressions were used as a starting point with only minor alterations made.

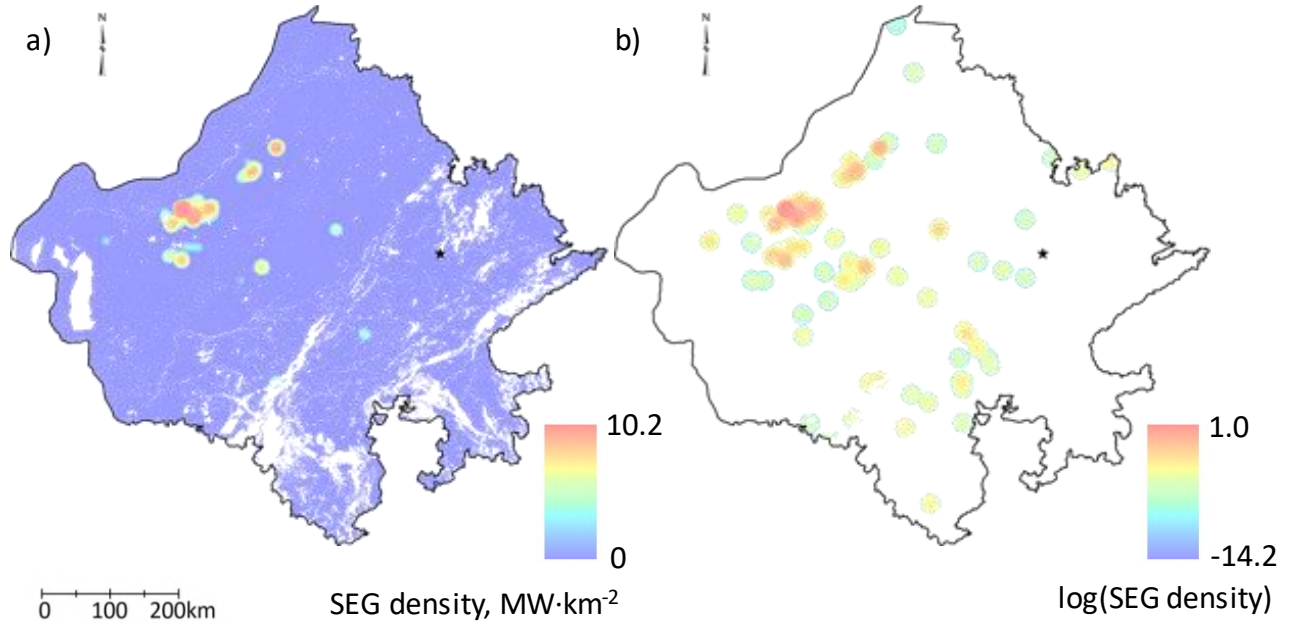


Figure 5: a) non-log and b) $\log(\text{SEG density})$, Rajasthan.

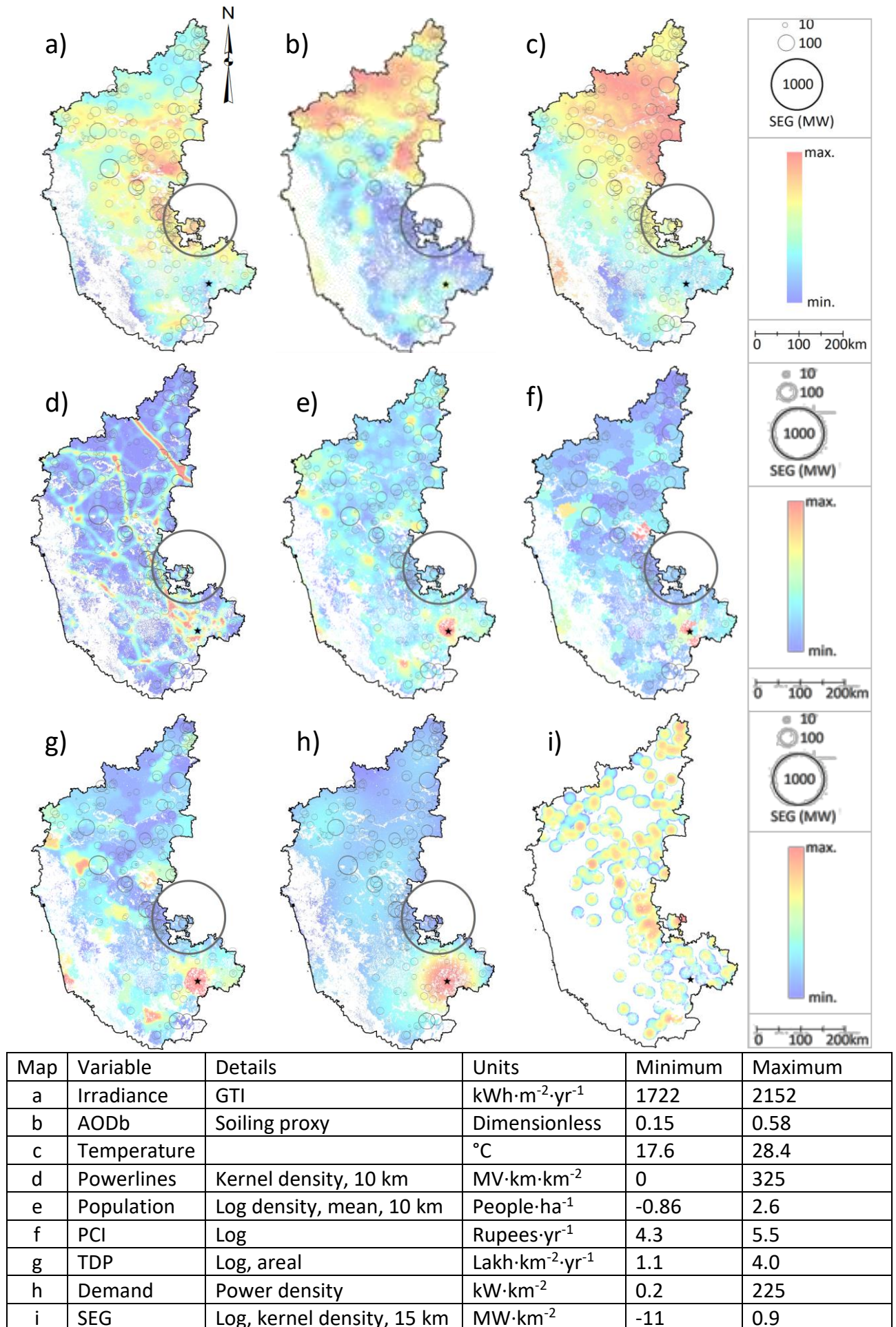
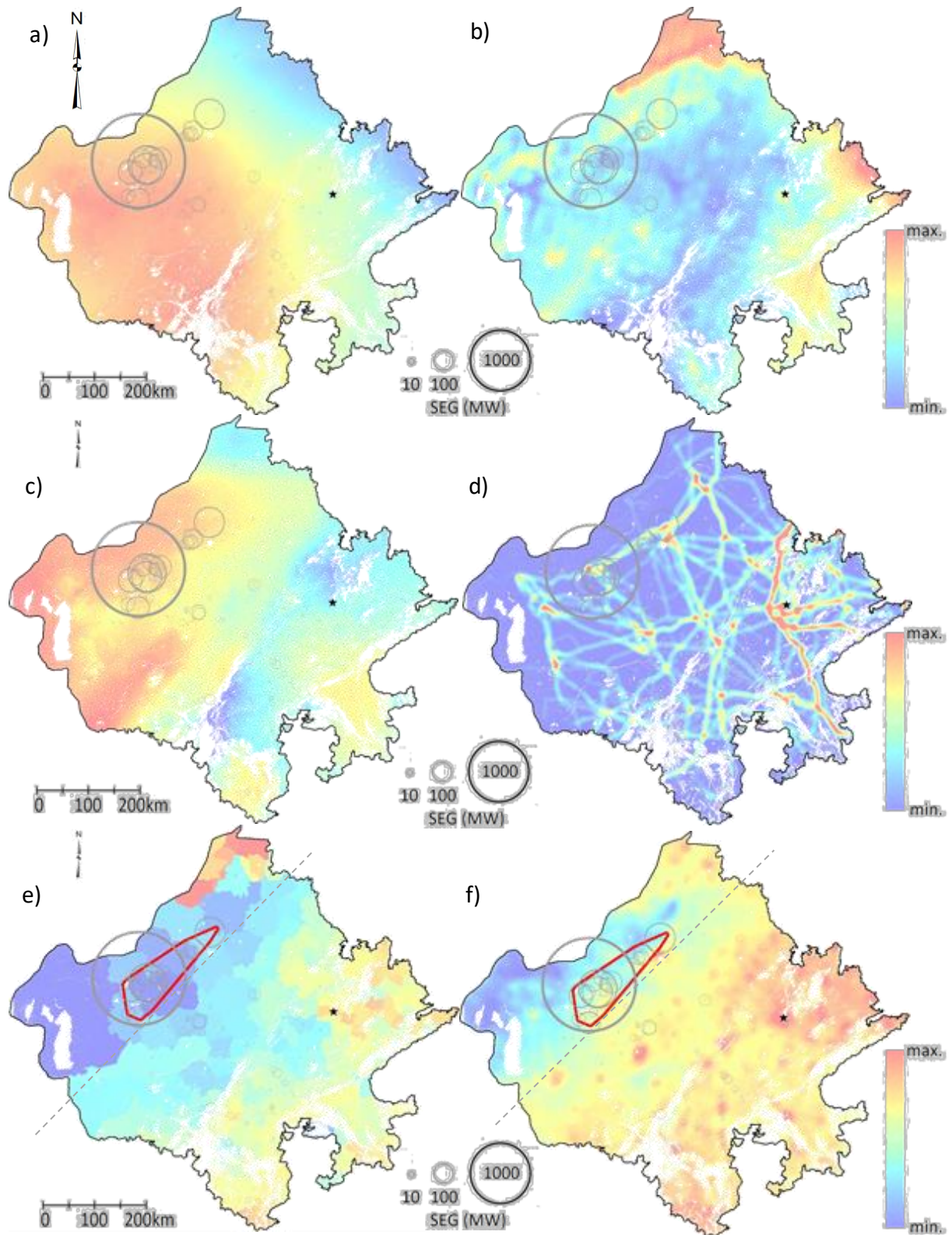


Figure 6: Karnataka input layers. State capital Bengaluru (★) dominates maps d-g. Note high northern Temperature and AODb.



Map	Variable	Details	Units	Minimum	Maximum
a	Irradiance	GTI	$\text{kWh} \cdot \text{m}^{-2} \cdot \text{yr}^{-1}$	1922	
b	AODb	Soiling proxy	Dimensionless	0.15	0.68
c	Temperature		$^{\circ}\text{C}$	21.0	28.3
d	Powerlines	Kernel density, 10 km	$\text{MV} \cdot \text{km} \cdot \text{km}^{-2}$	0	379
e	Population	Log density, mean, 10 km	$\text{People} \cdot \text{ha}^{-1}$	-1.6	1.8
f	GPB	Log, areal	$\text{rupees} \cdot \text{m}^{-2}$	-1.1	1.1

Figure 7: Independent variable layers, Rajasthan. Location of 90 % polygon west of transition axis in e, f suggests importance of land price proxies despite low Demand. Note similar GPB, Population and Demand patterns, suggesting Demand centroid lies east of centre. Roads layer not shown due to lack of visible pattern (Appendix 9.10).

3.3 Hypotheses

Basic hypothesis: the sign of the regression coefficient of each independent variable will agree with the influence assigned to it in [Table 2](#)/section [3.1.5](#) ($\beta_{\text{Irradiance}} > 0$, $\beta_{\text{AODb}} < 0$, etc.), with one modification: Analysis of SEG in Andhra Pradesh, bordering Karnataka, has found a negative relationship with population density (Flemming, 2020), indicating $\beta_{\text{population}} < 0$. Sign disagreement would imply inconsistency between siting practice and recommended criteria.

Extended hypothesis: each variable will be more coherent where more limiting, in the following sense:

For beneficial variables: having a more positive/less negative coefficient

For detrimental variables: having a more negative/less positive coefficient

For example, mean GTI is $2179 \text{ kWh}\cdot\text{m}^{-2}\cdot\text{yr}^{-1}$ Rajasthan and $2063 \text{ kWh}\cdot\text{m}^{-2}\cdot\text{yr}^{-1}$ in Karnataka, suggesting more positive/less negative⁶ $\beta_{\text{Irradiance}}$ in Karnataka.

3.4 Regression testing

All data layers were converted to rasters, resampled to 500 m resolution, snapped to a common grid, and restricted to the inclusion zone. Rasters were exported to ASCII files (one per variable) and combined into csv files (one per state). Each csv file contained one column per variable, one row per gridpoint. Regressions were run as backwards stepwise MLR in MatLab (MATLAB, 2017) (Appendix [9.11](#)) using SEG density or log(density) as the dependent variable. “Backwards stepwise” indicates that when a set of independent variables was entered the software automatically removed those that made the least statistically significant contribution until all $p < 0.05$.

A backwards-then-forwards guided ensemble approach was taken: in the backwards phase, large sets of independent variables were included. Variables were deprioritised for further analysis if they consistently met any of these criteria:

- 1) Weak statistical significance compared to other variables
- 2) Contradiction between sign of coefficient and relationship suggested in the literature
- 3) Variation in coefficient sign between regressions on the same region
- 4) Low effect size

The importance of (2) is this: this study aims to infer criteria used by SEG planners. If regression outputs cohere with criteria suggested in the literature, it may be proposed that these criteria have been used. If outputs for a given criterion consistently contradict the academic

⁶ Both forms must be considered: $\beta_{\text{Irradiance}}$ might be negative in both states, disconfirming the basic hypothesis, but less negative in Karnataka, corroborating the extended hypothesis.

consensus, this may indicate a confounding variable, or a disagreement between theoreticians and practitioners. Such contradictions will be discussed, but in neither case is there sufficient evidence for this study to propose a suitability index that qualitatively contradicts the consensus.

Relative effect size was indicated by β^* (Schroeder *et al.*, 1986) ([Equation 4](#)):

$$\beta_k^* = \frac{\sigma_{x_k}}{\sigma_y} \beta_k$$

Equation 4

Where β_k^* is the standardised regression coefficient of the k th independent variable x_k , σ_{x_k} and σ_y are the standard deviations of x_k and the dependent variable, and β_k is the regression coefficient of x_k . β_k^* is the change in y , measured in σ s, that results from a change in x_k of one σ (Schroeder *et al.*, 1986). It is unitless and takes account of the variability of dependent and independent variables, facilitating comparison (Nieminen *et al.*, 2013; Newman and Browner, 1991). Overall model coefficient of determination R^2 will be used to compare between similar regressions, but not as a standalone measure of regression utility: low correlation is a likely feature of highly disaggregated and variable spatial data (Wong, 2009).

The backwards phase produced a core of independent variables whose relationship to the dependent variable was stable, statistically significant, and consistent with the literature. During the forwards phase, deprioritised variables were reintroduced in conjunction with this core for further analysis. Different averaging areas for Population were also introduced. Social variables were excluded from the stable core to enable comparison between them in the forwards phase.

In Karnataka, where SEG is more evenly distributed than in Rajasthan, regional regressions were run on the northern and southern halves of the state. The forwards phase incorporated Moving Window Regression (MWR) to further evaluate spatial relationships (code: Appendix [9.12](#)).

In MWR, a regular grid is established ([Figure 8](#)). For each gridpoint, a specified regression is run on all data within a sub-region centred on that point. This enables assessment of spatial variation of the relationships under investigation (Fotheringham *et al.*, 2002: 42), as in GWR, but unlike GWR does not model coefficient variation as a function. Relationships between independent variables and their β^* values were analysed to evaluate the extended hypothesis that variables are more important when more limiting.

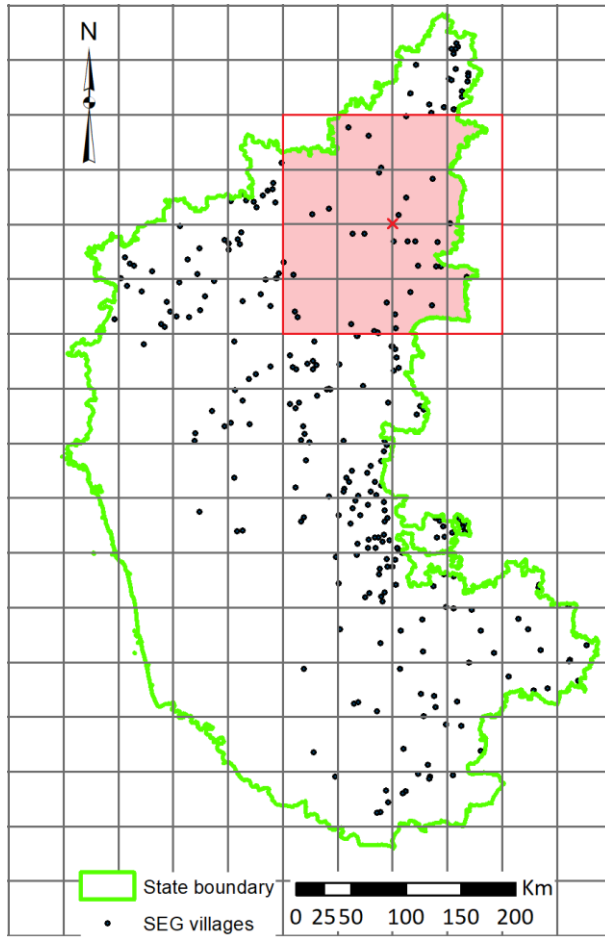


Figure 8: Example MWR grid. Spacing 50 km, bandwidth 200 km. Every vertex is a potential regression point. Data within the red square are used to calculate coefficients at **X**.

Due to scale differences, MWR results may disagree with higher level regressions. At state level it could be desirable to locate SEG in higher-income districts to access demand centres, but within districts to build in lower-income taluks to reduce costs. State-level regressions would then show positive β_{PCI} , while local-level (MWR) would show the opposite. As suitability indices may be used at either level, such disagreement will be a criterion for deprioritisation. MWR was not undertaken for Rajasthan because, due to SEG clustering, most windows would contain insufficient SEG data (see 90 % polygon, [Figure 4](#)).

In Karnataka, where northern and southern regression outputs were consistently different, these were combined as a Homogeneous Smooth Transition Regression (HSTR; see Teräsvirta, 1994). A general form of the MLR equation ([Equation 1](#)) can be modelled for the north:

$$y_N = \beta_{0_N} + \sum_{k=1}^K \beta_{k_N} x_k$$

Equation 5 ($_N$ indicates north)

And similarly for the south. For each coefficient $\beta_0 \dots \beta_K$, there is a north/south difference:

$$\Delta\beta_k = \beta_{k_N} - \beta_{k_S}$$

Equation 6 ($_s$ indicates south)

As a two-zone suitability map would introduce an unrealistic, abrupt transition, a three-zone map was created with northern, southern and transitional bands ([Figure 9](#)). There is a latitudinal distance ΔL between the northern and southern bands. To create a smooth transition, coefficients are modelled as varying steadily from south to north. Rate of change of β_k with latitude can be expressed as

$$\frac{\Delta\beta_k}{\Delta L}$$

Equation 7

Within the transitional band, coefficients can be modelled at any meridional distance l from the southern boundary with [Equation 8](#):

$$\beta_{kl} = \beta_{ks} + \frac{\Delta\beta_k}{\Delta L}l$$

Equation 8

The regression equation at any point in the transitional band becomes

$$y = \beta_{0s} + \frac{\Delta\beta_0}{\Delta L}l + \sum_{k=1}^K \left(\beta_{ks} + \frac{\Delta\beta_k}{\Delta L}l \right) x_k$$

Equation 9

The accuracy of the modelling assumption expressed in [Equation 7](#) was assessed with MWR analysis looking at how coefficients change through the transitional zone.

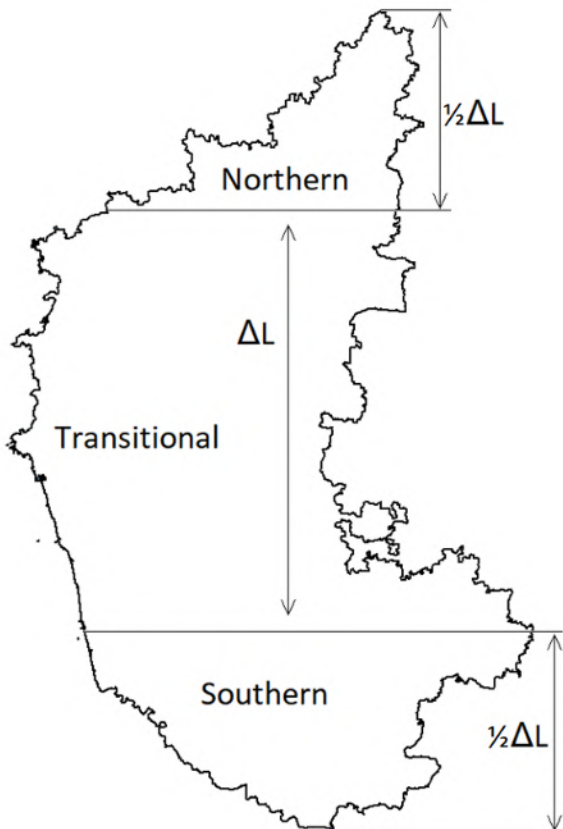


Figure 9: HSTR bands

These phases produced a set of regression equations to evaluate as suitability indices. Suitability maps were generated for each by entering independent variable raster layers $x_1...x_K$ and coefficients $\beta_0... \beta_K$ into global [Equation 1](#) and HSTR [Equation 9](#), in effect fitting trend surfaces ([Figure 10](#)). As the dependent variable modelled by non-log regressions is SEG density, suitability maps generated were of predicted SEG density at each point. For log runs the dependent variable is $\log(\text{SEG density})$, so trend surface values were exponentiated to produce predicted SEG density.

Mean and sum suitability scores were extracted for each SEG village for each mapped regression. Mean score for any village is its predicted mean SEG density, while the sum score is predicted total capacity. Correlations were checked between sum score and village capacity, and mean score and village SEG density. As Pearson's R is sensitive to outliers (Wilcox, 2017: 486), Spearman's R was also calculated.

Overall weighted averages were calculated. Mean scores were weighted by observed village SEG density, and sum scores by observed village capacity. Weighted averages for each regression were compared to the inclusion zone mean for that regression to produce a z-score, indicating how well the regression differentiates suitable from unsuitable areas. As MLR assumes that residuals (ϵ , [Equation 1](#)) are normally distributed and of constant variance (homoscedastic) (Pek *et al.*, 2018), these were assessed for mapped regressions. Heteroscedasticity may artificially reduce coefficient p values, but does not bias coefficients (Pek *et al.*, 2018), and is known to occur in studies of this nature (Balta-Ozkan *et al.*, 2015; Dharshing, 2017).

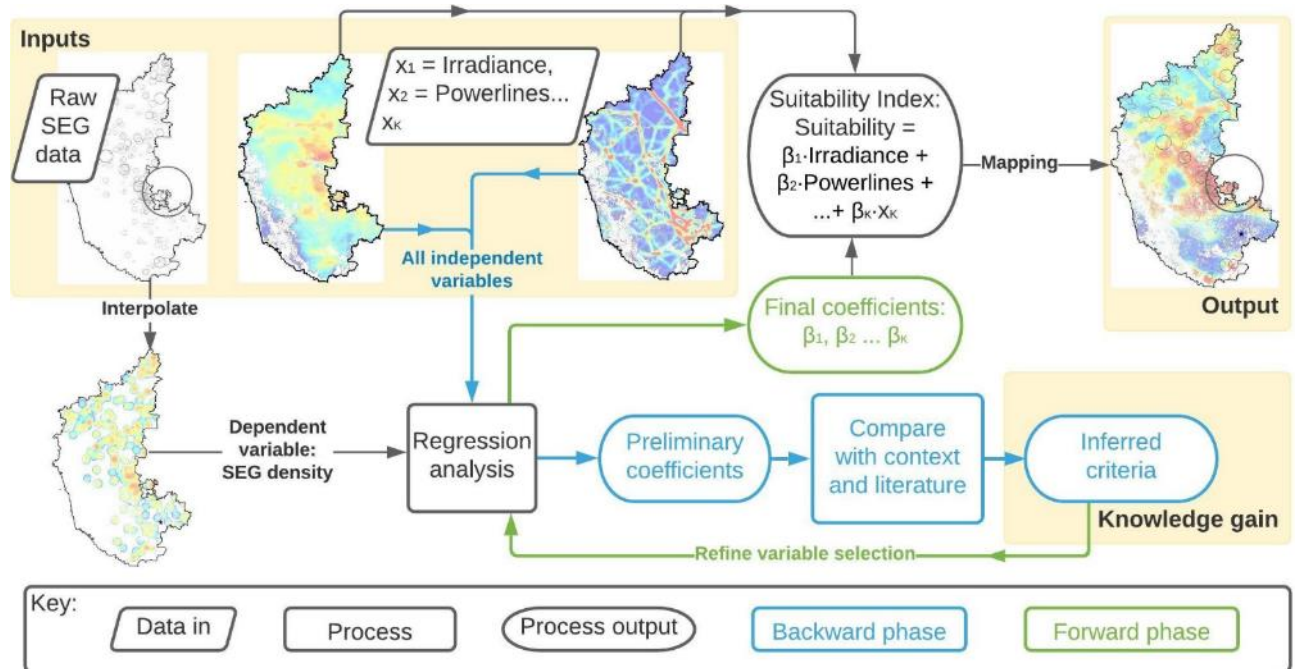


Figure 10: workflow. Note parallel with the fitting of a trend line to a graph: a regression is run on dependent and independent variables to produce a regression equation, then the independent variables entered into the equation to model the dependent variable. The difference here is the addition of the refinement process.

4 Results

As the regression refinement process was iterative, results will include some explanation where necessary to clarify choice of variables for further analysis.

4.1 Rajasthan

4.1.1 Backwards phase

17 backwards-phase regressions were run (Table 4). Slope and Powerlines were included in the core of stable variables because signs were consistently in agreement with the literature: β_{slope} was always negative, consistent with the understanding that steep slopes impede access and construction, while $\beta_{\text{powerlines}}$ was always positive, reflecting the grid access requirement. Inclusion of Powerlines made the greatest difference to goodness of fit: mean R^2 for regressions with Powerlines was 0.016, against 0.005 without.

Irradiance and AODb were deprioritised, due to removals and variable sign respectively. Temperature and Roads (i.e., distance to roads) were ruled out from further consideration because their coefficient signs contradicted the literature.

Table 4: Overview of backwards-phase results (mean values).

β^* is presented as absolute values to facilitate comparison. % Removal indicates the frequency with which a variable was found to be statistically insignificant ($p > 0.05$). $-\log p$ values are presented for ease of comparison between orders or magnitude. % Positive shows the frequency with which $\beta > 0$. Green (red) highlighting indicates 100 % consistence (disagreement) with expected sign. Note the small p values and large β^* for Powerlines and both social variables, compared to other variables.

Variable	Expected sign	Number of regressions	% Removal	$ \beta^* $, %	P value	$-\log(p \text{ value})$	% Positive
Irradiance	+	13	31	2.6	9.89E-03	2.0	100
AODb	-	15		1.7	3.81E-03	2.4	20
Temperature	-	7		5.7	1.06E-84	84	100
Slope	-	11		1.1	1.61E-07	6.8	0
Powerlines		8		11	0	NA	100
Roads	-	7	14	0.63	1.36E-02	1.9	83
GPB	-	5		7.7	0	NA	0
Population	-	6		7.0	4.06E-65	64	0





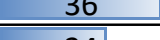

4.1.2 Forwards phase

With Temperature removed, $\beta_{\text{irradiance}}$ was positive in 100 % of regressions in both phases ($n = 20$, mean $\beta^* = 2.3$ %, mean p value = $5.1 \cdot 10^{-7}$). Irradiance was therefore included in all further regressions. β_{AODb} was negative in both phases for all regressions that included Population ($n = 22$, mean $\beta^* = -2.3$ %, mean p value = $8.8 \cdot 10^{-29}$). AODb was included in all further analysis of Population. With Temperature and Roads removed, β_{AODb} was negative for 90 % of regressions

across both phases ($n = 20$, mean $\beta^* = -1.7\%$, mean p value $= 8.5 \cdot 10^{-5}$). However, β_{AODb} was positive for 33 % of such regressions that had GPB rather than Population as the social variable ($n = 3$, mean $\beta^* = 0.0\%$, p value $= 2.7 \cdot 10^{-4}$). The regression using Irradiance, Slope, Powerlines and GPB was selected for mapping and evaluation (regression R-GPB). Full equations for all mapped regressions: Appendix [9.13](#).

The forwards phase included analysis of four Population variables. Population was averaged at 10, 20 and 30 km (Pop₁₀, Pop₂₀, Pop₃₀), and over a combination of annular regions out to 35 km. Regressions were run on each together with AODb and Slope plus (1) Irradiance, (2) Powerlines, or (3) both. The combined radial-annular approach generated higher $|\beta^*|$, with consistent signs and lower p values than other variables ([Table 5](#), cf. [Table 4](#)). The negative-positive-negative pattern was consistent across all regressions run on annular Population ($n = 8$), including log regressions ($n = 2$). The regression using Irradiance, AODb, Slope, Powerlines and radial-annular Population was selected for mapping and evaluation (regression R-Pop).

Table 5: Overview of forwards-phase regression results for Population

	Radius, km	$ \beta^* $, %	P value	-Log(p value)	% Positive
Separate radial	10	 9.4	0	NA	0
	20	 7.8	0	NA	0
	30	 7.7	0	NA	0
Combined radial-annular	15	 21	3.1E-266	266	0
	15-25	 36	0	NA	100
	25-35	 24	1.1E-275	275	0

4.1.3 Mapping and Evaluation

[Figure 11](#) shows suitability maps for non-log R-GPB and R-Pop. Starting from these combinations for log regressions showed the same independent variables to be consistent with the literature, with the addition of Roads ([Figure 12](#)). These maps are compared with a control regression (Irradiance, Slope and Powerlines) in [Table 6](#). Residuals for the non-log regressions were severely non-normal; those for log regressions were within acceptable bounds. Residuals were heteroscedastic in both cases, but less severely so for log regressions (Appendix [9.14](#)).

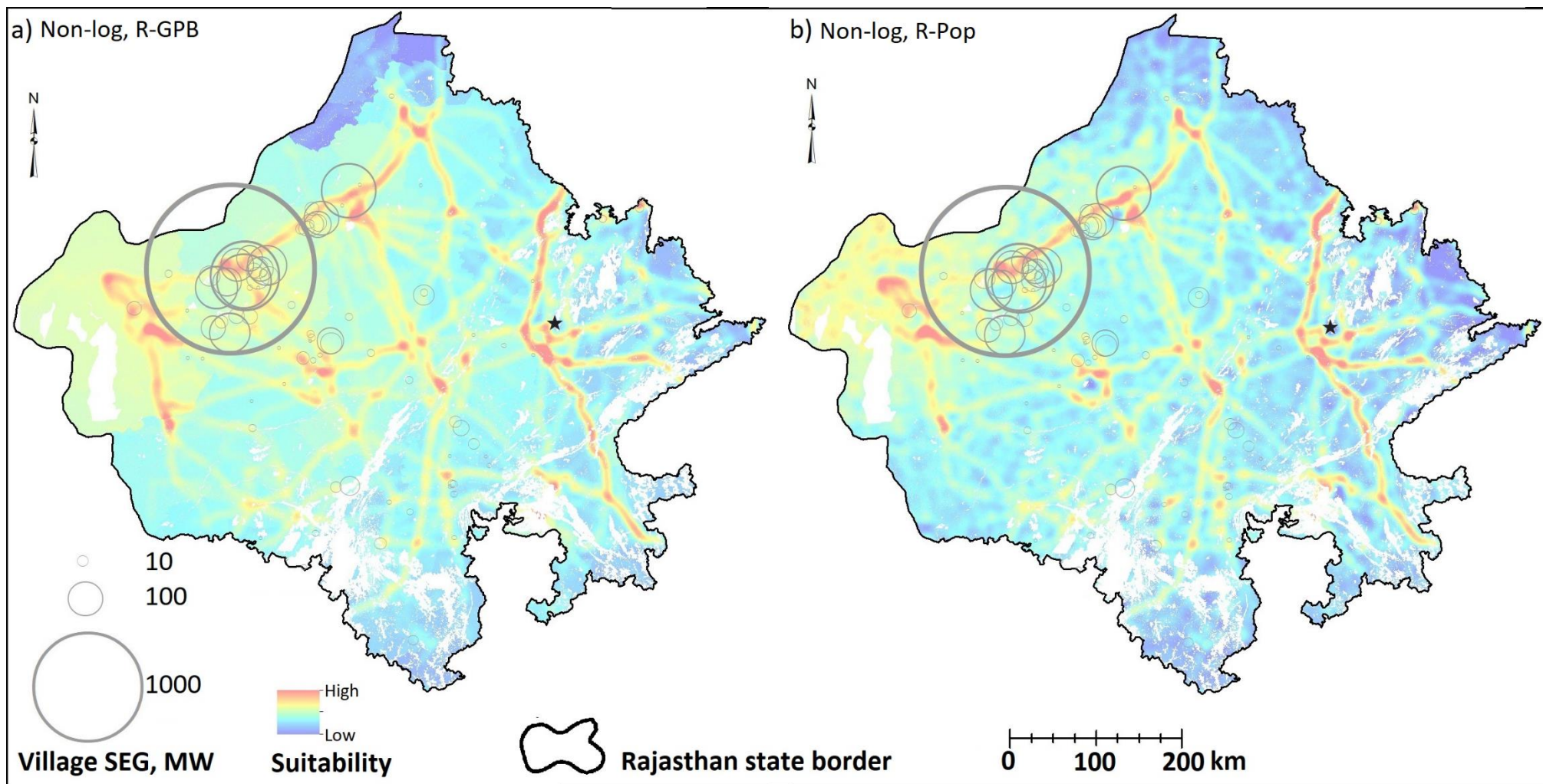


Figure 11: Suitability maps based on non-log regressions using Irradiance, Slope, and Powerlines. (a) (non-log R-GPB) also includes GPB; (b) (non-log R-Pop) includes AODb and annular Population. Boundaries between taluks with different GPB are visible in (a), as are powerline corridors on both maps. Exclusion zone left blank.

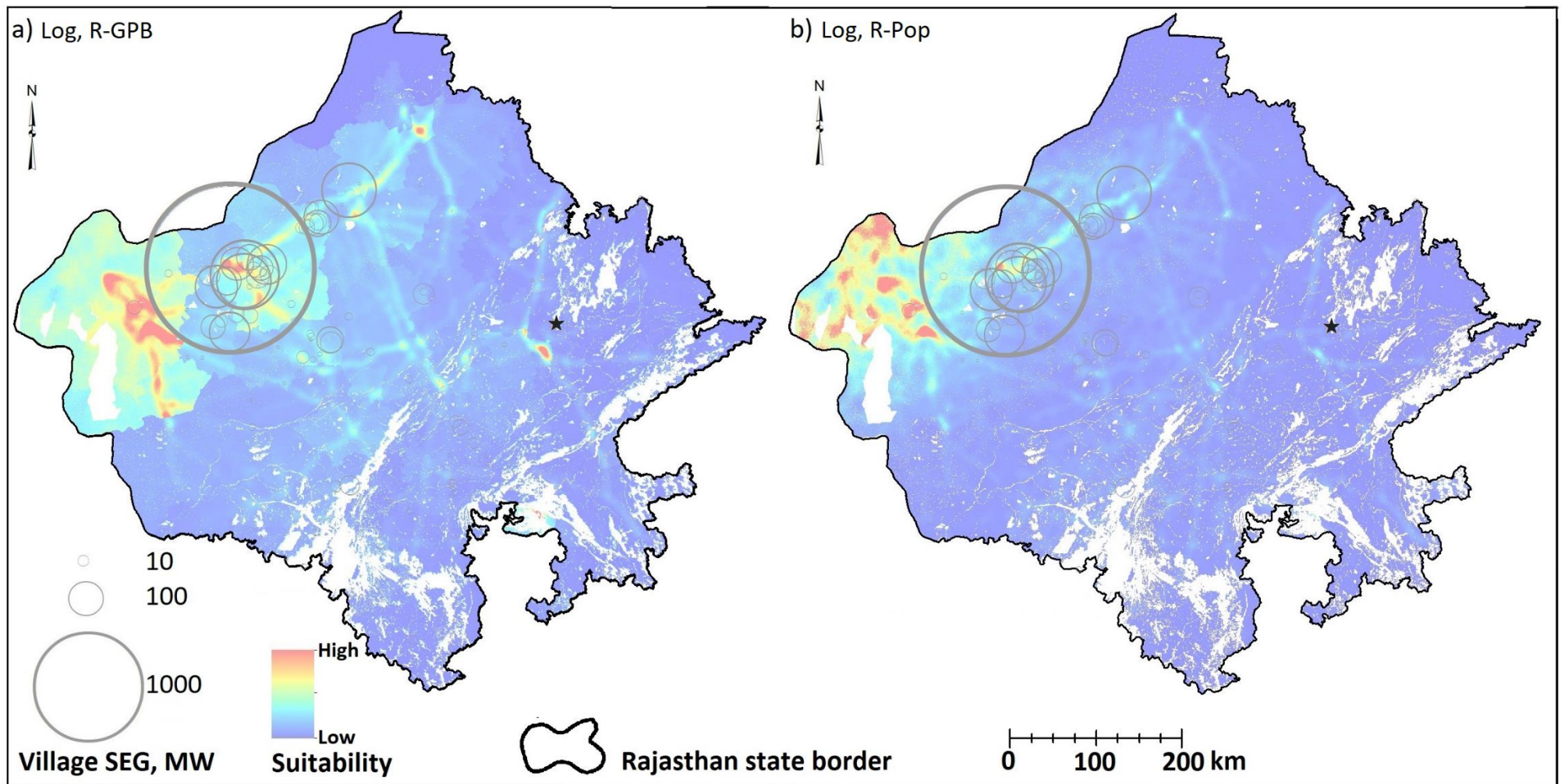


Figure 12: Suitability maps based on log regressions using Irradiance, Slope, Powerlines, AODb and Roads. (a) (log R-GPB) includes GPB; (b) (log R-Pop) includes annular Population. The outline of low-GPB Jaisalmer district in the extreme west is visible in (a).

Table 6: Rajasthan regression evaluation statistics.

* p value ≤ 0.001 ; † $p \leq 0.01$; ‡ $p \leq 0.1$. P values of z -scores not evaluated. Highest value in each row in **bold**, lowest value underlined. More green (red) indicates higher (lower) values. Areal weighting leads to more stringent evaluation of all statistics. Log regressions produce some of the highest z -scores and correlation strengths, but also some of the lowest. R-GPB outperforms R-Pop for log regressions only.

Comparison	Statistic	Non-log			Log	
		Control	R-Pop	R-GPB	R-Pop	R-GPB
Sum score, capacity (MW)	Z score	<u>176</u>	218	205	180	314
	Pearson	0.46 *	0.42 *	0.39 *	<u>0.31</u> †	0.40 *
	Spearman	<u>0.28</u> †	0.32 †	0.36 *	0.37 *	0.40 *
Mean score, density (MW·km ⁻²)	Z score	59	55	52	<u>30</u>	75
	Pearson	0.03	0.07	<u>-0.01</u>	0.06	0.08
	Spearman	<u>-0.07</u>	-0.05	-0.03	-0.04	-0.01
Model R ²		<u>0.011</u>	0.018	0.016	0.121	0.130

4.2 Karnataka

4.2.1 Backwards phase

76 non-log and 63 log regressions were run (Table 7, Table 8). Population variable used was pop₁₀.

Table 7: Overview of backward-phase non-log regression results (mean values).

Bold values indicate important points of difference from log results. Highlighting as before. Note that $\theta_{population}$ is always positive in North, but never in Global or South. Both are consistent with the literature, though the modified basic hypothesis predicts $\theta_{pop} < 0$. All mean p values < 0.01 .

Category	Variable	Expected sign	Global			North			South		
			n	$ \beta^* $, %	% +	n	$ \beta^* $, %	% +	n	$ \beta^* $, %	% +
Climate	Irradiance	+	22	8.2	100	20	2.3	0	20	11	100
	AODb	-	26	7.0	0	26	8.1	0	24	3.1	21
	Temperature	-	9	1.2	100	6	6.4	100	7	1.5	14
Terrain	Slope	-	25	2.4	0	25	2.4	0	23	1.9	0
IS	Powerlines	+	25	4.6	100	25	3.7	100	23	5.9	100
Social	Population	+, -	8	7.9	0	10	2.3	100	7	12	0
	PCI	-	9	3.3	0	8	12	0	8	0.0	63
	TDP	-	8	4.0	13	8	3.5	0	9	2.4	33
	Demand	+	15	5.8	0	12	3.6	100	13	6.9	0

Table 8: Overview of backwards-phase log regression results. Data bars are standardised across both tables, with generally higher $|\beta^*|$ values in the log regressions. All mean p values <0.01 .

Category	Variable	Expected sign	Global			North			South		
			n	$ \beta^* $, %	% +	n	$ \beta^* $, %	% +	n	$ \beta^* $, %	% +
Climate	Irradiance	+	17	6.3	100	17	0.6	82	17	14	100
	AODb	-	21	9.8	0	21	14	0	21	12	0
	Temperature	-	5	13	100	5	9.3	100	5	3.6	100
Terrain	Slope	-	21	3.5	0	21	2.4	0	21	3.7	0
IS	Powerlines	+	19	0.3	63	19	5.1	100	19	4.3	5
Social	Population	+,-	7	13	0	7	0.2	57	7	21	0
	PCI	-	9	7.9	0	9	9	0	9	2.25	22
	TDP	-	9	13	0	9	1.7	0	9	15	0
	Demand	+	12	14	0	12	2.7	0	12	11.6	0

Core variables for non-log regressions were selected as follows:

Global: Coefficient signs of Irradiance, AODb, Slope, and Powerlines were consistent with the literature and therefore included. Temperature and Demand were not, and were rejected. TDP and PCI tended to have the anticipated sign but lower $|\beta^*|$ and were deprioritised relative to Population.

North: AODb, Slope, Demand and Powerlines were included in the core of stable variables.

Temperature and Irradiance were rejected. Coefficient signs of TDP, PCI and Population were consistent with the literature; PCI was prioritised due to larger $|\beta^*|$ and because Population was removed once.

South: As for Global, except that AODb was deprioritised due to inconsistent sign, low $|\beta^*|$ value and two removals, and Temperature was retained for further analysis.

Core variables were the same for log regressions, with the following exceptions:

Global: Powerlines showed inconsistent sign and low $|\beta^*|$, and was removed three times. TDP and Population had equal mean $|\beta^*|$. Core variables: Irradiance, AODb, Slope. Population and TDP retained.

North: The Demand pattern was reversed, and $\beta_{\text{Population}}$ was inconsistent. Although $\beta_{\text{Irradiance}}$ was frequently positive when included, unlike the negative $\beta_{\text{Irradiance}}$ values in the non-log regressions, it was removed in 8 out of 17 regressions. Core variables: AODb, Slope, and Powerlines. PCI prioritised.

South: AODb and TDP were wholly consistent with the literature, and Temperature was wholly contradictory. Powerlines was generally contradictory, and removed once. Core variables: Irradiance, AODb, and Slope. Population prioritised.

4.2.2 Forwards phase

4.2.2.1 Non-log

Global: Pop₁₀ – Pop₄₀ were analysed. Coefficients were consistently negative, with slightly higher $|\beta^*|$ and overall R^2 values for Pop₂₀ ($|\beta^*_{\text{pop20}}| = 9.9\%$; $R^2 = 0.023$). Regression selected for mapping: Irradiance, AODb, Slope, Powerlines, Pop₂₀.

North: Assessment of PCI, TDP and Pop₁₀ to Pop₄₀ confirmed PCI as the most predictive variable, with mean $R^2 = 0.019$ for regressions with PCI ($n = 2$) and 0.006 without ($n = 6$). Candidate variables selected for further analysis: AODb, Slope, Powerlines, Demand, PCI.

South: All radial population variables were consistently negative, with $|\beta^*| > 11\%$ and R^2 between 0.032 and 0.034. Pop₂₀ was selected as it gave the highest $|\beta^*|$, 13 %. Results for AODb and Temperature were inconclusive. Candidate variables: Irradiance, AODb, Temperature, Slope, Powerlines, Pop₂₀.

4.2.2.2 Log

Global: In combination with the core, TDP gave weaker overall correlation ($R^2 = 0.050$) than Population variables (mean $R^2 = 0.060$). Pop₃₀ had the highest $|\beta^*|$ (21 %, against 19 % for TDP). Selected for mapping: Irradiance, AODb, Slope, Pop₃₀.

North: PCI gave the highest R^2 and $|\beta^*|$ values (0.025 and 8.5 %, against maxima of 0.018 and 1.1 % for regressions using population statistics). Candidate variables: AODb, Slope, Powerlines, PCI.

South: All Population variables gave overall $R^2 \geq 0.115$ and $|\beta^*_{\text{pop}}| > 24\%$. Pop₃₀ gave the highest values, with $R^2 = 0.142$ and $|\beta^*_{\text{pop}}| = 29\%$. Candidate variables: Irradiance, AODb, Slope, Pop₃₀.

4.2.3 Moving Window Regression

4.2.3.1 Non-log MWR

MWR with all candidate variables shows the local effect of each ([Figure 13](#)). Each datapoint shows β^* for the variable in a 200x200 km window, plotted against the mean value of that variable in that window. Datapoints are coloured according to latitude. A horizontal colour gradient indicates a meridional trend in the variable, a vertical one indicates a meridional trend in effect size. Effect size may vary approximately smoothly with latitude (Irradiance, Temperature, Population; AODb somewhat), be approximately constant with latitude (Slope, Powerlines, PCI), or exist in two regimes (Demand). Correlation strengths were slightly weaker for β than β^* ([Appendix 9.15](#)).

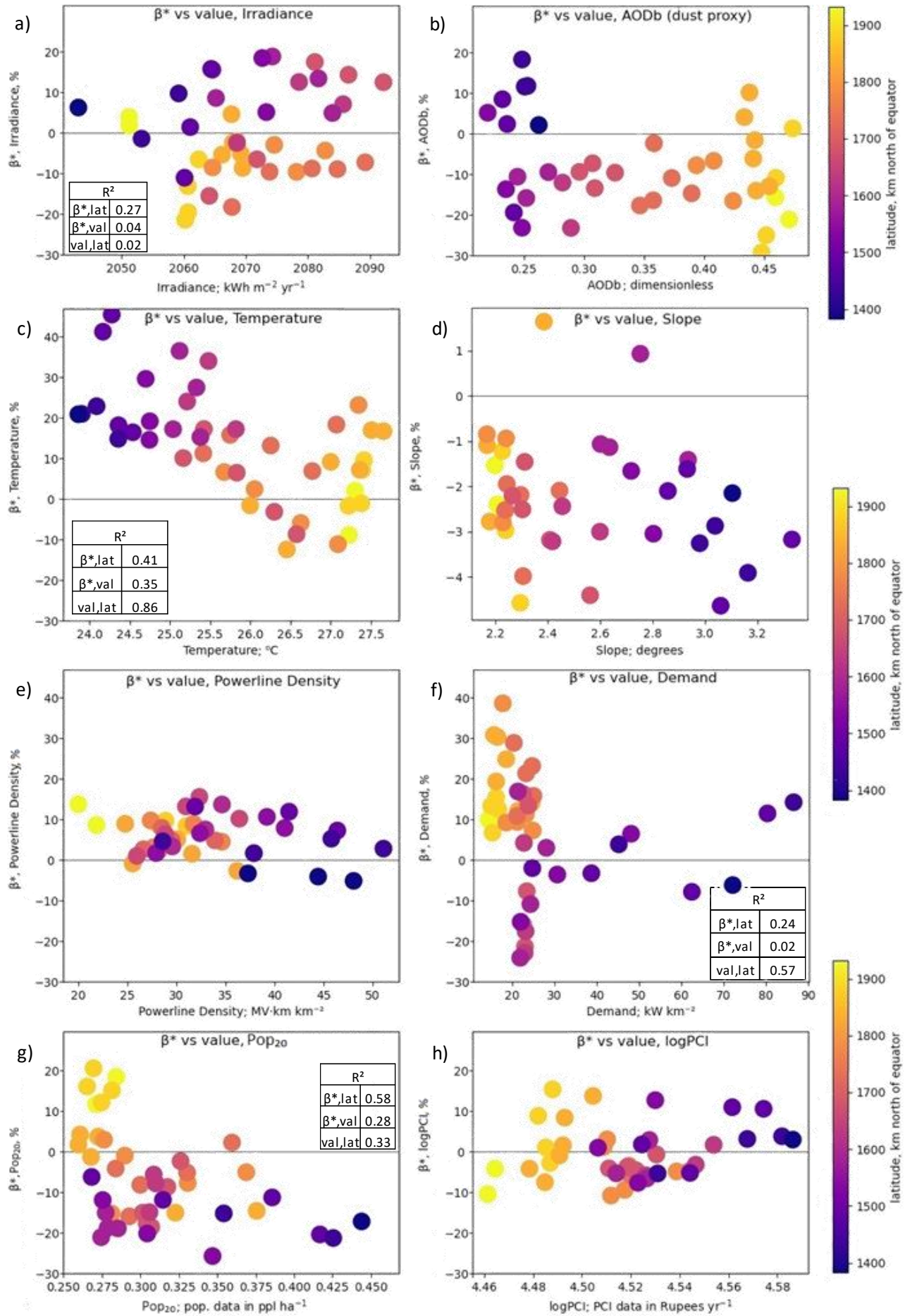


Figure 13: β^* for candidate variables in MWR, non-log. Spacing 50 km, bw 200 km. Y axes visually comparable except for Slope. Insets: correlation strengths. Not shown if < 0.1 for both β^* correlations. Note greater correlation of $\beta^*_{\text{Irradiance}}$ with latitude than with Irradiance values, and similarly for Pop₂₀.

β^*_{pop20} is well correlated with latitude (Figure 13), suggesting Pop_{20} could have been used for both regions in the HSTR. However, in combination with only northern candidate variables, β^*_{pop20} was negative, contradicting the local pattern. Northern β^*_{PCI} was strongly positive at regional level (Table 7) but inconclusive locally (Figure 13h). MWR with northern candidate variables and bandwidth 300 km yielded all northern $\beta^*_{\text{PCI}} < 0$ (Figure 14). Variables were selected for mapping as shown in Table 9.

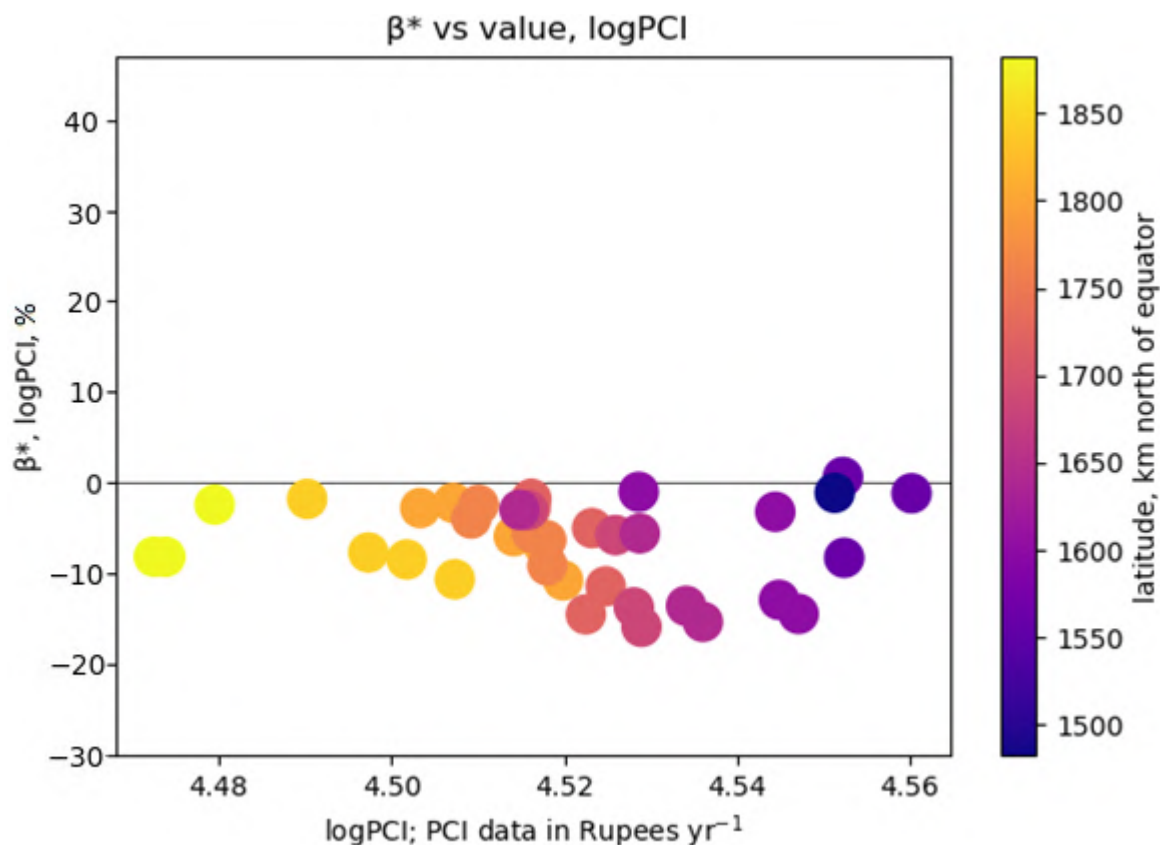


Figure 14: MWR output, PCI. Northern variables only. Spacing 40 km, bandwidth 300 km. Non-log.

Table 9: Relationships between each variable and non-log SEG density

Relationships indicated by β^* in regional and MWR regressions. Decisions about where to retain each for HSTR and mapping are also shown. “- -” indicates β^* negative in both regions but more than twice as large in this one.

Category	Variable	Relationship found				Retained for HSTR in	Reason
		Regional		MWR			
		N	S	N	S		
Climate	Irradiance	-	+	-	+	S	N contradicts literature
	AODb	--	-	-	+-	N, S	Consistent trend; both broadly consistent with lit.
	Temperature	+	-	+-	+	Neither	MWR trend contradicts regional trend
Terrain	Slope	-	-	-	-	N, S	Consistent trend; both consistent with literature
IS	Powerlines	+	+	+	+	N, S	Consistent trend; both consistent with literature
Social	Population	+	-	+	-	S	N shows stronger relationship with PCI
	PCI	-	0	-	-	N	S shows stronger relationship with Population
	Demand	+	-	+	-	N	S contradicts literature

4.2.3.2 Log MWR

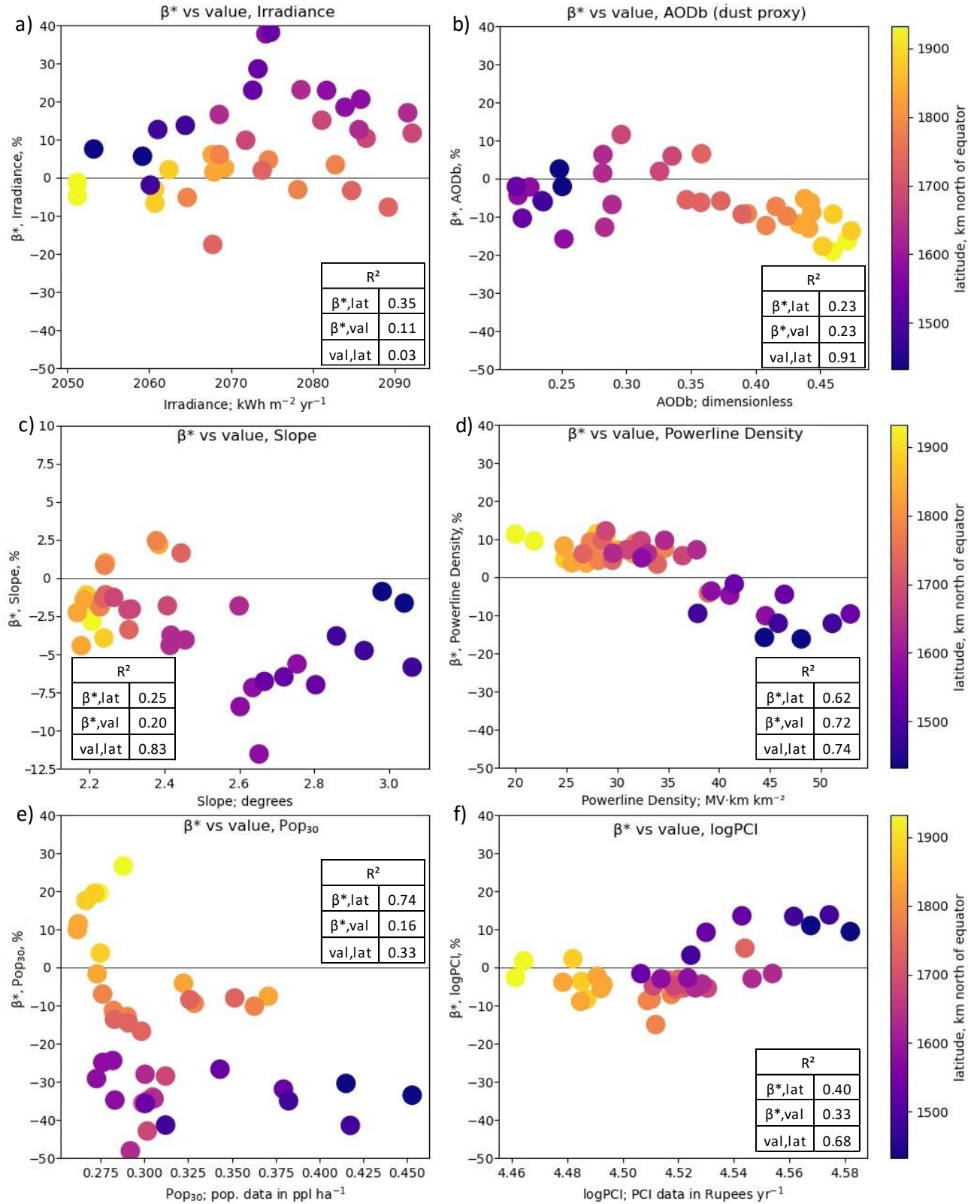


Figure 15: β^* , log MWR. Spacing 50 km, bandwidth 200 km. Note generally stronger correlations than for non-log data. Some value-latitude correlations change because only data within 15 km of SEG villages is included.

Log MWR (Figure 15) shows trends largely as for non-log, with certain exceptions. The β^* _{powerlines} trend is more pronounced, and outputs now suggest a negative relationship between PCI and SEG density at low PCI but a positive one at high PCI (Figure 15f). To test this unexpected trend

($R^2 = 0.40$), MWR was rerun with bandwidth 300 km and northern variables only ([Figure 16](#)). As before this reduced the incidence of $\beta^* > 0$, though southern outliers remain. Northern β^*_{PCI} is generally negative, as expected, so PCI was retained for northern regressions ([Table 10](#)).

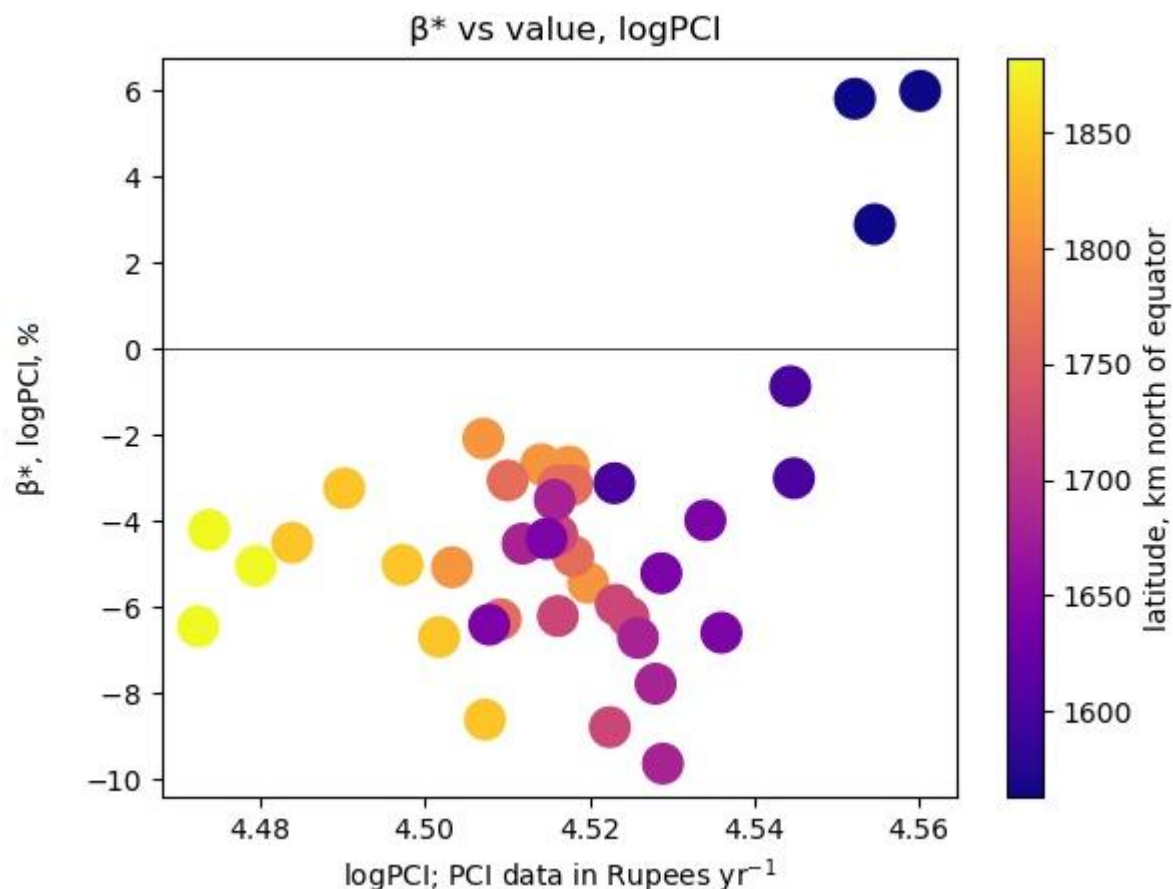


Figure 16: MWR output: β^*_{PCI} . Northern candidate variables. Spacing 40 km, bandwidth 300 km.

Table 10: Selection of variables for log HSTR mapping and evaluation.

Category	Variable	Relationship found				Retained for HSTR in	Reason
		Regional		MWR			
		N	S	N	S		
Climate	Irradiance	-	+	-	+	S	N contradicts literature
	AODb	-	-	-	-	N, S	Consistent trend; both consistent with literature
Terrain	Slope	-	-	-	-	N, S	Consistent trend; both consistent with literature
IS	Powerlines	+	-	+	-	N	S contradicts literature
Social	Population	0	-	+ -	-	S	N shows little relationship with Population
	PCI	- -	-	-	+ -	N	S shows stronger relationship with Population

4.2.4 Mapping and evaluation

Suitability maps generated from these regressions are presented in [Figure 17](#) and [Figure 18](#), and evaluated in [Table 11](#). A non-log control was run on Irradiance, Slope and Powerlines. As in Rajasthan, non-log residuals were severely non-normal, log residuals were within acceptable bounds, and all were heteroscedastic with log regressions less severely so ([Appendix 9.14](#)). The implications of this will be addressed in the Discussion.

Table 11: Karnataka evaluation statistics.

* p value ≤ 0.001 ; † $p \leq 0.01$; ‡ $p \leq 0.1$. P values of z -scores not evaluated. Highest value in each row in **bold**, lowest value underlined. Log regressions tend to outperform non-log, HSTR tends to outperform global, and all outperform the control. Log HSTR also yields the most significant areal Spearman correlation ($p = 0.21$).

Comparison	Statistic	Non-log			Log	
		Control	Global	HSTR	Global	HSTR
Sum score, capacity (MW)	Z score, %	<u>55</u>	97	107	100	130
	Pearson	<u>0.14</u> ‡	0.23 *	0.24 *	0.24 *	0.25 *
	Spearman	<u>0.19</u> †	0.22 *	0.26 *	0.26 *	0.27 *
Mean score, density (MW·km ⁻²)	Z score, %	<u>25</u>	59	66	47	69
	Pearson	<u>0.00</u>	0.15 ‡	0.16 †	0.21 *	0.18 †
	Spearman	<u>-0.08</u>	0.04	0.07	0.03	0.08
Model R ² (Shown separately for N and S components of HSTR)		0.009	0.023	S: 0.033 N: 0.019	0.060	S: 0.141 N: 0.025

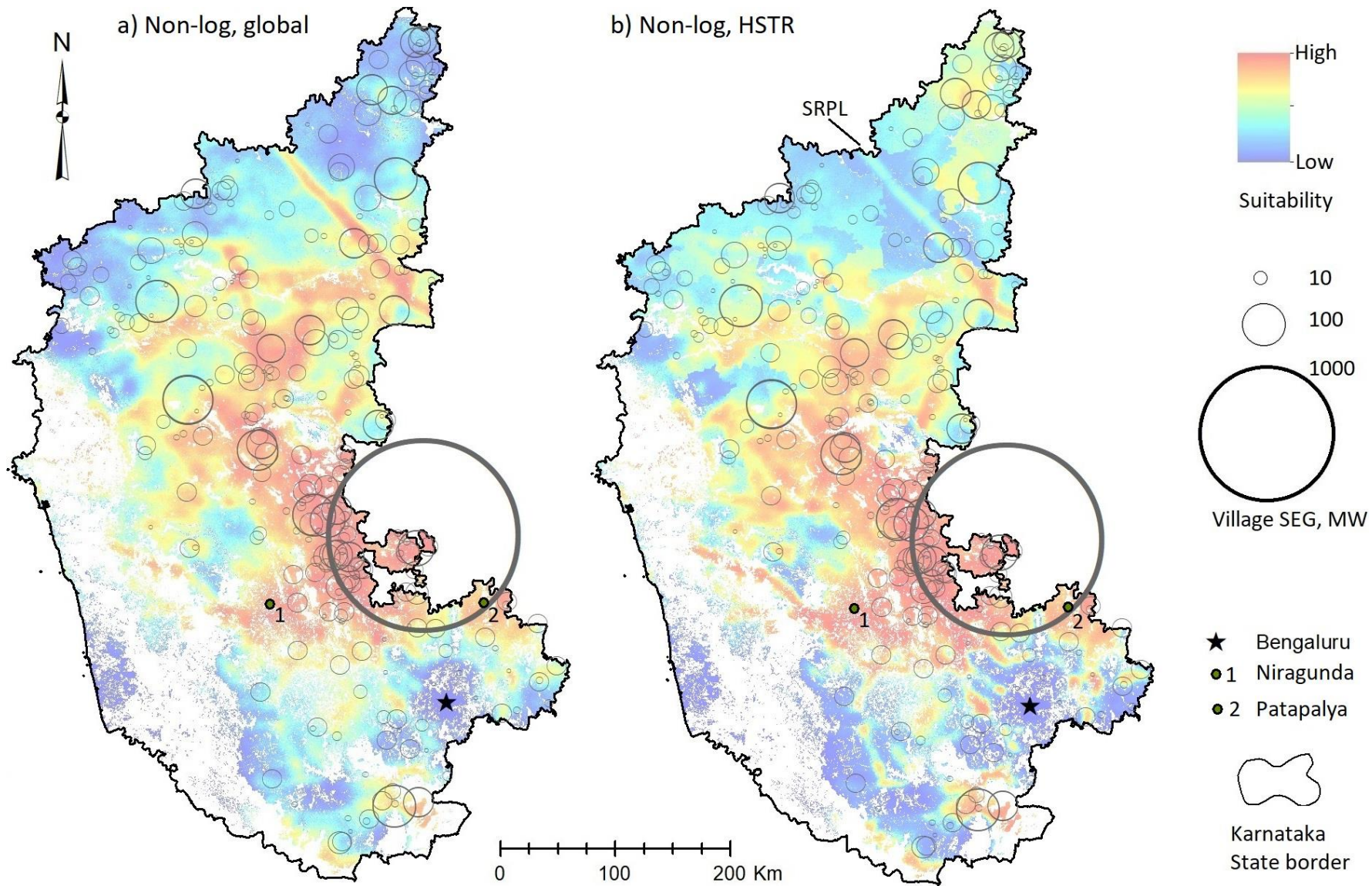


Figure 17: Suitability maps generated from (a) global and (b) HSTR non-log regressions. (a) uses Irradiance, AODb, Slope, Powerlines, and Pop₂₀. (b) uses these in the south but AODb, Slope, Powerlines, Demand and PCI in the north. (a) highlights the length of the 765 kV Solapur-Raichur powerline (SRPL) as highly suitable, but HSTR mutes this in high-PCI taluks. Colour scales are equalised within each map to enhance contrast

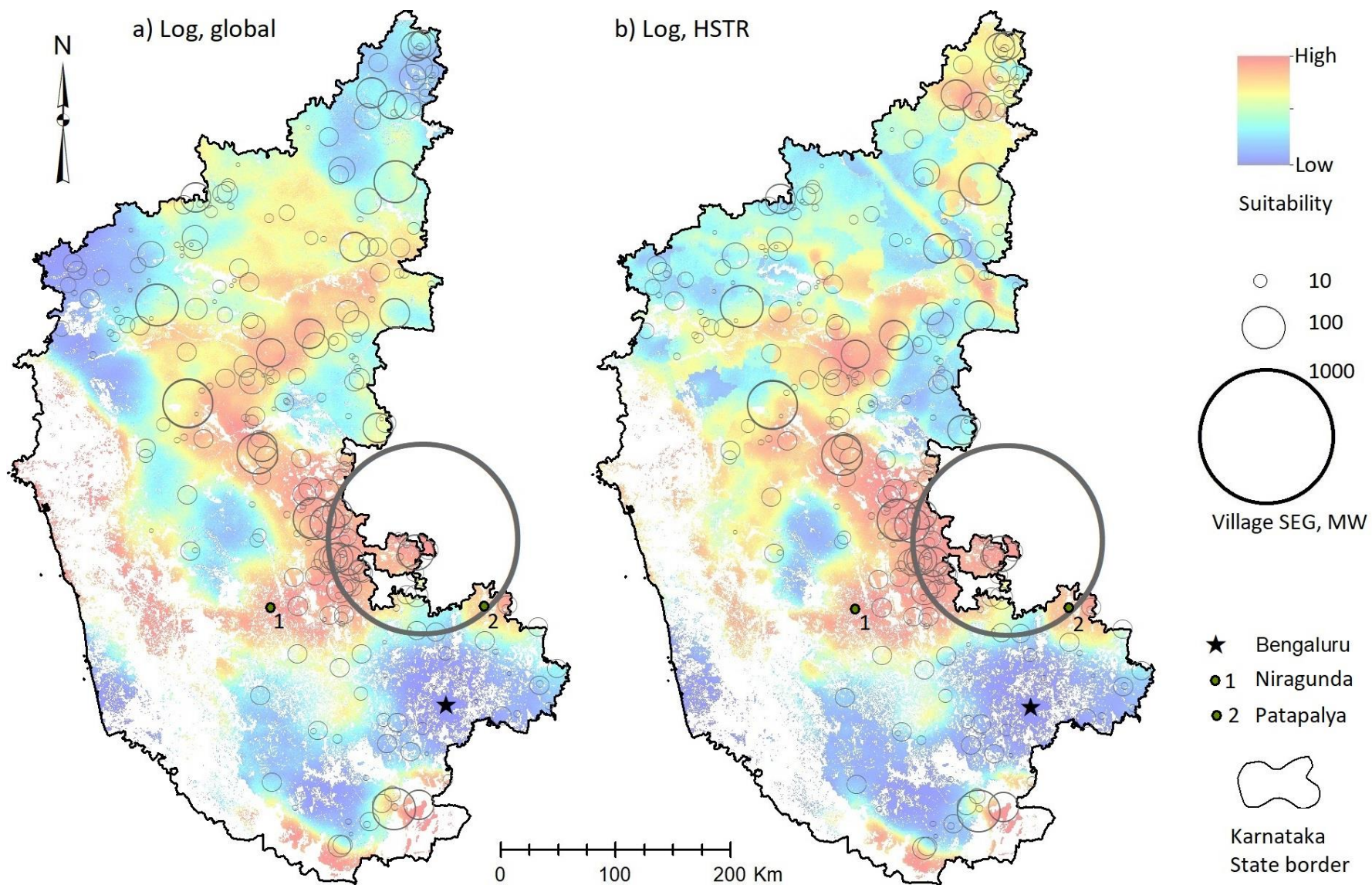


Figure 18: Suitability maps generated from global and HSTR log regressions. (a) uses Irradiance, AODb, Slope, Pop_{30} . (b) uses these in the south but AODb, Slope, Powerlines, PCI in the north. (b), by using PCI in the north and tailored coefficients, correctly picks out low-PCI, high-SEG areas in the far north. All maps indicate underexploited high-suitability land around Niragunda and Patapalya, though the Niragunda area is fragmented by excluded land.

5 Discussion

5.1 Weight analysis

This section will discuss each variable evaluated, stating whether the expected relationship was found (basic hypothesis) and whether it was more coherent⁷ where the variable was more limiting (extended hypothesis). Residuals evaluation has shown that the log regressions, which used log-transformed SEG density as the dependent variable, are more statistically valid. Outputs for Karnataka therefore refer to log regressions, unless otherwise stated. For Rajasthan the blindness of log regressions to 87 % of the dataset is considered more problematic, and discussion will focus on outputs from non-log regressions, despite the loss of confidence that this entails.

5.1.1 Climate

5.1.1.1 Irradiance

Irradiance is of low importance in Rajasthan, high importance in South Karnataka, and negative importance in North Karnataka, as measured by β^* . Contradiction of the basic hypothesis in North Karnataka may be explained by intercorrelations: in the north, Irradiance is positively correlated with land price proxies (Population, TDP, PCI), and weakly or neutrally correlated with Slope and AODb⁸. In the south, Irradiance is neutrally or negatively correlated with the proxies, and more strongly anticorrelated with slope and AODb ([Figure 19](#)).

⁷ See earlier definition

⁸ MLR models relationships with multiple variables, but only linearly, so intercorrelation between independent variables still presents difficulties if they have non-linear relationships with the dependent variable.

North		SEG		Climate			Ter'n	IS	Social				South	
		log(SEG)	SEG	Irradiance	AODb	Temperature	Slope	Powerlines	Demand	PCI	TDP	Population		
Social	Population	0.00	0.02	0.14	-0.06	-0.04	0.04	0.20	0.34	0.25	0.59		Population	Social
	TDP	-0.03	0.00	0.10	-0.21	-0.15	0.05	0.17	0.37	0.64		0.78	TDP	
	PCI	-0.06	-0.09	0.09	-0.23	-0.22	0.04	0.23	0.07		0.72	0.40	PCI	
	Demand	-0.01	0.04	0.33	-0.22	-0.14	0.04	0.06		0.36	0.60	0.54	Demand	
IS	Powerlines	0.04	0.03	0.15	0.00	0.00	-0.01		0.36	0.16	0.28	0.29	Powerlines	IS
Ter'n	Slope	-0.02	-0.02	0.00	-0.07	-0.01		-0.07	0.00	0.11	0.02	-0.07	Slope	Ter'n
Climate	Temperature	0.02	0.02	-0.09	0.49		-0.07	-0.04	-0.31	-0.07	-0.15	-0.14	Temperature	Climate
	AODb	-0.13	-0.06	-0.09		0.08	0.11	-0.04	-0.04	0.29	0.35	0.33	AODb	
	Irradiance	-0.04	-0.01		-0.33	0.29	-0.30	0.10	-0.13	-0.35	-0.21	0.00	Irradiance	
	SEG	0.65		0.13	-0.08	0.03	-0.05	0.04	-0.07	-0.06	-0.09	-0.10	SEG	
SEG	log(SEG)		0.32	0.26	-0.13	0.12	-0.07	-0.11	-0.27	-0.17	-0.30	-0.34	log(SEG)	SEG
		log(SEG)	SEG	Irradiance	AODb	Temperature	Slope	Powerlines	Demand	PCI	TDP	Population		
		SEG		Climate		Ter'n	IS		Social					

Figure 19: Intercorrelation of independent variables, Karnataka (Pearson's R). Values for North (South) Karnataka are above (below) the central diagonal, arranged symmetrically. Coefficients referred to in the text in **bold**. Note strong intercorrelation of social and infrastructure variables (top right corner), particularly in the south. IS: Infrastructure. SEG: SEG density. Ter'n: Terrain. Continuous scale from dark green (strong positive correlation) through yellow (no correlation) to red (anticorrelation).

This indicates more land in the south that has high Irradiance in conjunction with desirable levels of these other variables. The suitability maps reflect this: the south has extensive areas of very high/low suitability (deep reds/blues), with concentrations and near-absences of SEG respectively, while the north has intermediate scores (pale blues/yellows) and smaller, sparser SEG ([Figure 17](#), [Figure 18](#)). This may also be reflected in lower model R^2 values, which are generally higher in the south ([Figure 20](#)): with fewer areas of very high or very low suitability, it is harder to establish clear siting criteria.

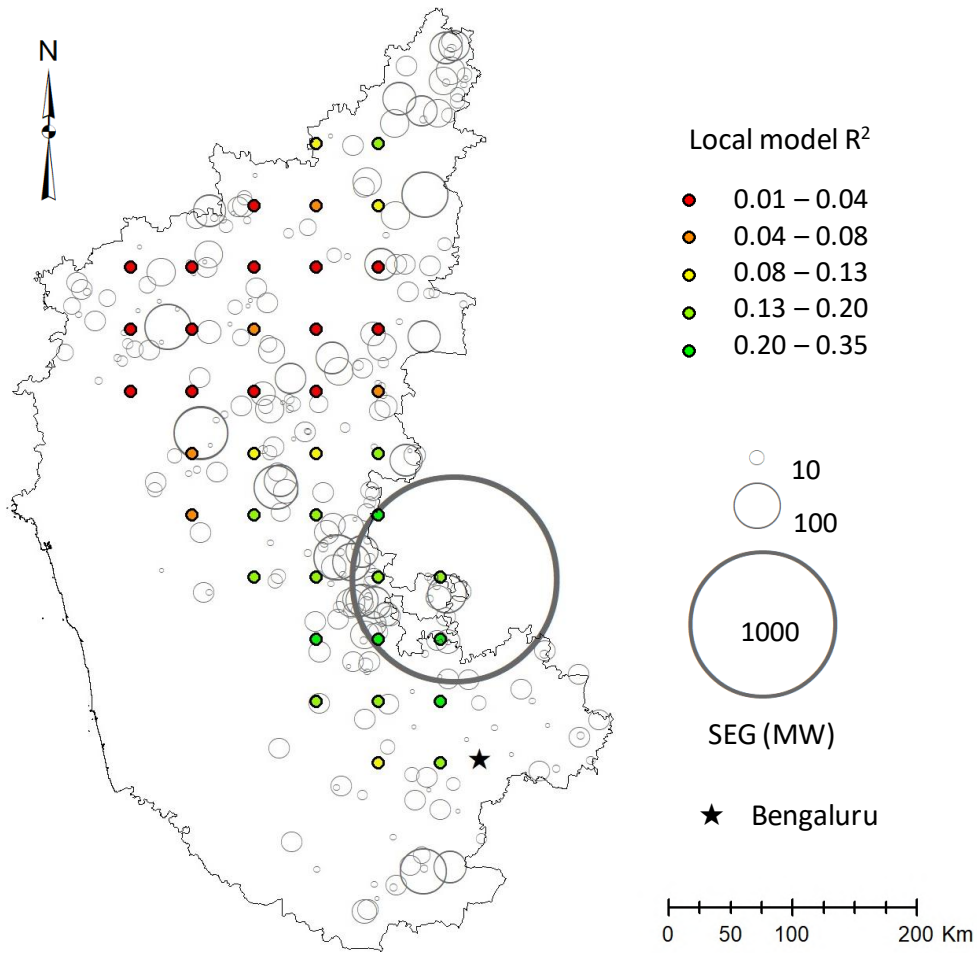


Figure 20: R^2 values from local log regressions. MWR run on all candidate variables for north and south, spacing = 50 km, bandwidth = 200 km. Low coverage in west and far south due to insufficient SEG data in windows. Note higher correlation in SEG-dense south/central region.

As Irradiance is beneficial, it is more limiting where lower, so the extended hypothesis predicts more positive $\beta^*_{\text{irradiance}}$ in such areas. At regional and inter-state level this occurs ([Table 12](#)) though MWR shows a stronger correlation of $\beta^*_{\text{irradiance}}$ with latitude than with Irradiance ($R^2 = 0.35$ vs 0.11). This suggests that other spatially varying characteristics influence its importance, arguing against a general relationship. Weights in previous studies are generally higher than those found here, but vary widely, as evidenced by the large inter-quartile range (IQR, a non-parametric measure of dispersion) ([Figure 21](#)).

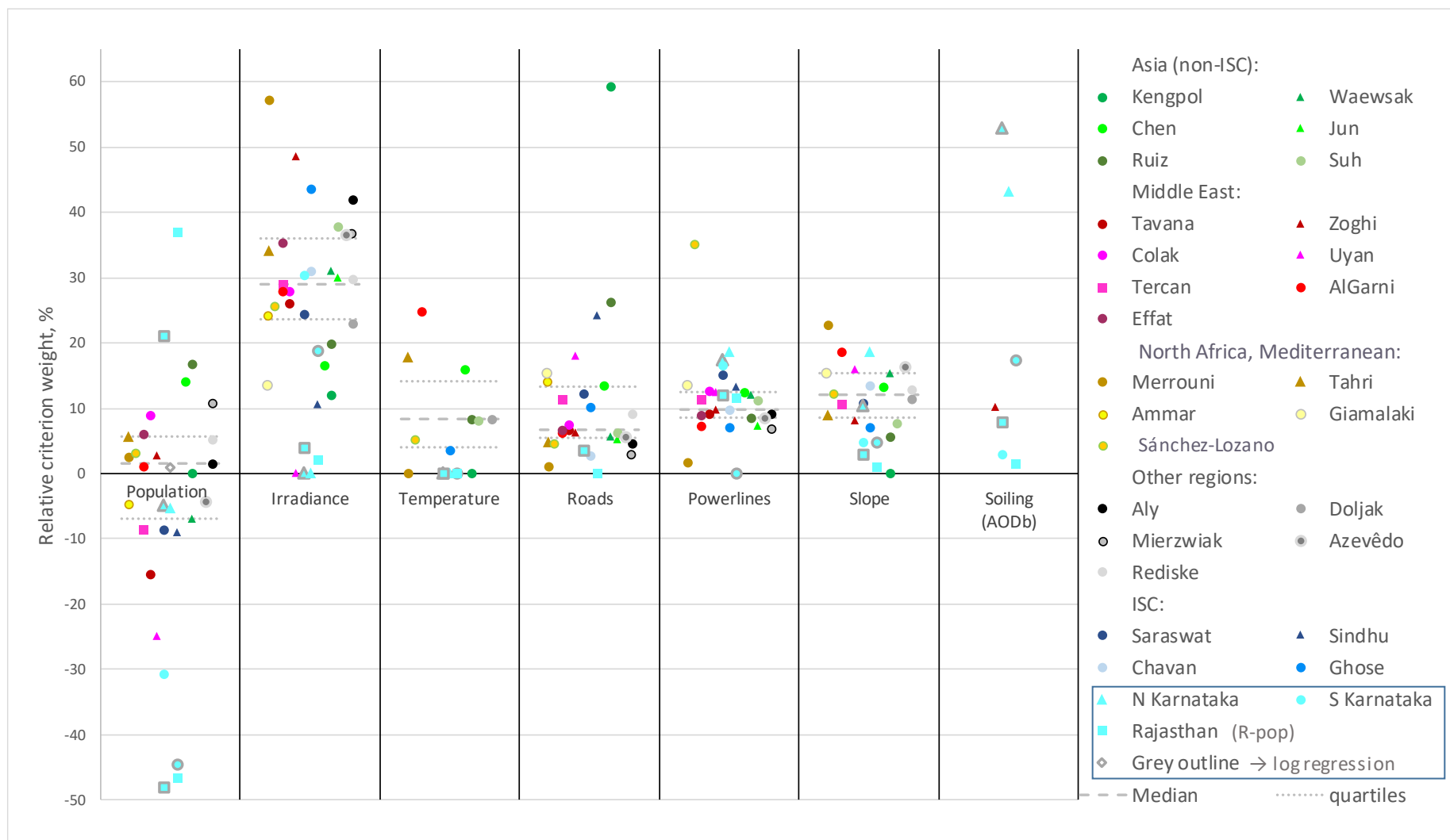


Figure 21: Comparison with previous studies (identified by lead author; see references). To facilitate comparison, weights are standardised (see Appendix 9.16) and given as positive values, except to differentiate between positive and negative weighting of Population. The difference between quartiles is the IQR. Regressions plotted are those mapped (hence positive and negative Population components in R-Pop), except in north Karnataka, where Population variables replace PCI to increase comparability. Regression equations: Appendix 9.13.

Table 12: Comparison of hypothesised and observed inter-regional β^* differences

For each variable, cells are green if β^* is more coherent in the region where the variable is more limiting, corroborating the extended hypothesis, and red otherwise. † indicates an inter-state comparison that would give a different result if log regression β^* were used for Rajasthan. β^* values given are means of all backwards phase regressions. Population comparisons, and especially those between states, should be treated with caution as averaging distances had regionally different impacts.

Category	Variable	Units	Expected influence	a) Comparing North and South Karnataka (log regressions)						b) Comparing Karnataka (log), Rajasthan (non-log; (log))					
				Mean value		More limiting in	β^* , %, mean		β^* more coherent in	Mean value		More limiting in	β^* , %, mean		β^* more coherent in
				North	South		North	South		Ktk	Rj		Ktk	Rj	
Climate	Irradiance	kWh m ⁻² yr ⁻¹	+	2067	2059	S	0.6	14	S	2063	2179	Ktk	6.3	2.6 (1.7)	Ktk
	AODb	None	-	0.43	0.26	N	-14	-12	N	0.34	0.36	Rj	-10	-1.7 (-8.8)	Ktk
	Temperature	°C	-	26.7	24.7	N	9.3	3.6	S	25.7	26.5	Rj	13	5.7 (15.8)	Rj†
Terrain	Slope	°	-	2.3	3.0	S	-2.4	-3.7	S	2.6	2.4	Ktk	-3.5	-1.1 (-2.5)	Ktk
IS	Powerlines	MV km km ⁻²	+	27.8	36.4	N	5.1	-4.3	N	32.1	30.8	Rj	-0.30	10.6 (13.4)	Rj
Social	Pop ₁₀	ppl ha ⁻¹	-	0.43	0.48	S	0.17	-21	S	0.45	0.031	Ktk	-13	-7	Ktk
	PCI	Rupees yr ⁻¹	-	4.5	4.6	S	-8.6	-2.3	N	Not assessed in Rajasthan					
	TDP	Lakhs yr ⁻¹ m ⁻²	-	1.9	2.0	S	-1.7	-15	S						
	Demand	kW km ⁻²	+	19.6	43.8	N	-2.7	-12	N						

5.1.1.2 AODb

β_{AODb} is consistently negative, corroborating the basic hypothesis. This suggests dust deposition may be an important criterion, though importance varies regionally ([Figure 21](#)). $|\beta^*|_{\text{AODb}}$ is much greater in Karnataka than Rajasthan, despite lower mean AODb ([Table 12](#)). This contradicts the extended hypothesis. Rajasthan's low values for $|\beta^*|_{\text{AODb}}$ are not obviously explained by intercorrelation, as AODb is negatively correlated with Irradiance and positively correlated with Population and GPB, suggesting that high-AODb land will also be low-Irradiance and high-price ([Figure 22](#)). Rajasthan log regressions give much higher $|\beta^*|_{\text{AODb}}$, comparable to Karnataka ([Table 12](#)). This is a more plausible result, but the scale of the difference underscores the lower confidence in Rajasthan outputs. Within Karnataka the extended hypothesis is corroborated, with higher $|\beta^*|_{\text{AODb}}$ in the high-AODb north. MWR weakly reflects this trend ([Figure 13b](#), [Figure 15b](#)).

Rajasthan		SEG		Climate			Ter'n	IS	Soc.
		log(SEG)	SEG	Irradiance	AODb	Temperature	Slope	Powerlines	GPB
Social	Population	-0.26	-0.07	-0.39	0.10	-0.68	-0.11	0.31	0.75
	GPB	-0.33	-0.08	-0.55	0.35	-0.65	-0.08	0.16	
IS	Powerlines	0.08	0.08	-0.14	0.00	-0.26	-0.10		
Ter'n	Slope	-0.03	-0.01	0.05	-0.06	0.07			
Climate	Temperature	0.29	0.06	0.46	0.02				
	AODb	-0.11	-0.03	-0.59					
	Irradiance	0.18	0.05						
SEG	SEG	0.39							

Figure 22: Intercorrelations, Rajasthan. Temperature is strongly anticorrelated with social variables, and positively correlated with Irradiance. The opposite pattern holds for AODb, though correlations with social variables are weaker. Colour scale as before.

5.1.1.3 Temperature

$\beta_{\text{temperature}}$ is consistently positive, contrary to expectations, suggesting a confounding variable. The obvious candidate is Irradiance, but $\beta_{\text{Temperature}}$ is positive in North Karnataka regressions despite anticorrelation of Temperature with Irradiance ($R = -0.09$) and Irradiance being frequently removed as insignificant. Temperature cannot simply be a proxy for Irradiance.

Temperature is also anticorrelated with all land price proxies, sometimes strongly ($R = -0.68$ between Temperature and Pop_{10} in Rajasthan), suggesting interrelationships not captured here. These results contradict the basic hypothesis. Consequently the weight for Temperature is 0 in all indices generated. [Figure 21](#) shows 0 values in two previous studies (Merrouni *et al.*, 2018; Kengpol *et al.*, 2017), and that Temperature is among the least frequently considered variables.

The extended hypothesis suggests that $\beta_{\text{Temperature}}$ would be lower where Temperature is higher. This is corroborated by non-log MWR, and inter-state evidence: in Karnataka, local-level $\beta^*_{\text{Temperature}}$ decreases as temperatures rise ([Figure 13c](#)), while Rajasthan, which has higher mean temperatures, has less positive $\beta^*_{\text{Temperature}}$ ([Table 12b](#)). This is not the case if log outputs are used in Rajasthan. The extended hypothesis is also contradicted at regional level: mean Temperature is higher in North than South Karnataka, yet $\beta^*_{\text{Temperature}}$ is more positive in the north ([Table 12a](#)).

5.1.2 Terrain

5.1.2.1 Slope

β_{slope} is negative across all regressions in all regions, corroborating the basic hypothesis. $|\beta^*|$ is larger in regions where Slope is larger ([Table 12](#)), supporting the extended hypothesis. MWR corroborates this at local level, though correlations are weak ([Figure 13d](#), [Figure 15c](#)). The range of weights is consistent with previous studies ([Figure 21](#)) and suggests that slope is of low but sustained importance.

5.1.3 Infrastructure

5.1.3.1 Powerlines

$\beta_{\text{powerlines}}$ is positive in all regressions, supporting the basic hypothesis, except global and southern log regressions in Karnataka. This discrepancy may be due to the log transform: areas with 0 SEG density may also have low powerline density, but this information is unavailable to log regressions.

This suggests that powerline density is negatively related to SEG density near southern SEG villages. This is borne out by intercorrelations. This negative relationship may be due to positive correlations between Powerlines and land price proxies, which are stronger in the south than the north ([Figure 19](#)). This, together with higher southern means for these proxies, suggests that the most powerline-dense southern land may be prohibitively expensive. This is consistent with the high powerline density and land price proxies around state capital Bengaluru in the south ([Figure 6, d-g](#)).

The north-south trend, reflected in the MWR, shows that $\beta^*_{\text{Powerlines}}$ is more positive where Powerlines is more limiting, as does inter-state comparison ([Table 12](#)), corroborating the extended hypothesis. Weights are consistent with previous studies ([Figure 21](#)) and suggest that powerline density is an important criterion.

5.1.3.2 Roads

In Rajasthan, results from log and non-log regressions for Roads are contradictory. In Karnataka, data proved unreliable. This precludes evaluation of the hypotheses.

5.1.4 Social

5.1.4.1 Population

Magnitude and range of $\beta^*_{\text{Population}}$ are generally greater than previous studies' weights. This suggests that Population is an important but complex criterion in the study area. The IQR for Population in previous studies is likewise higher for any other variable assessed ([Figure 21](#)). Results in Karnataka suggest a positive role for Population at low values and a negative role at high values, consistent with the extended hypothesis, but $\beta^*_{\text{population}}$ is more strongly correlated with latitude ($R^2 = 0.74$) than with Population ($R^2 = 0.16$), suggesting a role for other factors in the variation of $\beta^*_{\text{Population}}$. High $|\beta^*|$ for annular Population in Rajasthan, with negative weights for Pop_{15} and Pop_{25-35} but positive weight for Pop_{15-25} , are suggestive but inconclusive. Mixed weight signs disconfirm the basic hypothesis, but are consistent with bidirectional impacts of population density found by (Balta-Ozkan *et al.*, 2021) on domestic PV adoption.

5.1.4.2 Economic measures

For Rajasthan, GPB has the second highest mean $|\beta^*|$ in the backwards phase, and is negatively related to SEG in all regressions, while in Karnataka β_{TDP} and β_{PCI} are negative in 100 % of global and 95 % of regional backwards-phase regressions, all corroborating the basic hypothesis. TDP is higher in the south, implying higher land prices and greater limitation on SEG. The more negative southern β^*_{TDP} is thus consistent with the extended hypothesis, although the size of the difference appears disproportionate, suggesting other factors may be involved. PCI is also higher in the south than the north, so the more negative northern β^*_{PCI} values are inconsistent with the extended hypothesis. Both TDP and PCI are more influential than Irradiance, as measured by $|\beta^*|$, suggesting that economic criteria are important in the study area and that TDP and PCI are effective proxies.

5.1.4.3 Demand

Although distribution of SEG in Rajasthan appears anticorrelated with Demand ([Figure 4](#)), data was insufficient to assess this. For Karnataka, 100 % of non-log regressions show β_{demand} positive in the north and negative in the south. This may be linked to correlation between Demand and land price: mean R for Demand with Population, TDP and PCI is 0.26 in the north and 0.50 in the south, suggesting greater northern availability of high-Demand, low-land-price areas. All log regressions give negative β_{demand} . Results overall contradict the basic hypothesis. This merits further investigation as proximity to demand centres is frequently given as a reason to site SEG near population centres (e.g. Aly *et al.*, 2017; Merrouni *et al.*, 2018; Zoghi *et al.* 2017).

North/south comparison within Karnataka is consistent with the extended hypothesis, but the low correlation in MWR analysis between Demand and β^*_{demand} ($R^2 = 0.02$, non-log, [Figure 13f](#)) suggests other factors are involved.

5.2 Method Evaluation

For Rajasthan capacity-weighted outputs, z-scores are high, and correlations between suitability score and capacity are statistically significant ($p \leq 0.01$) ([Table 6](#)). This suggests successful identification of areas with large capacity as more suitable. Density-weighted correlations are not significant ($p \geq 0.1$), and in some cases are negative, suggesting unsuccessful differentiation between highly and moderately suitable areas. The more complex regressions do not consistently outperform the control, despite greater statistical validity of log regressions. MWR or log analysis was challenging due to the highly clustered SEG.

This clustering may also contribute to the greater heteroscedasticity of the Rajasthan models' residuals: the regression software attempts to fit a curve composed of linear multiples of smoothly varying independent variables, but the dependent variable has extensive 0 or near-0 values interrupted by isolated peaks around individual UMSPs. The regression models these rare values less well, with the result that $|\epsilon|$, and thus error variance, are larger when values are larger. Given that the precise location of UMSPs is not determined by easily quantified measures, this may be unavoidable. Despite these shortcomings, the western area highlighted by all regressions as most extensively suitable ([Figure 11](#), [Figure 12](#)) agrees with previous research in Rajasthan ([Figure 23](#)). This suggests that residual heteroscedasticity is not overly problematic for the method used here, which is consistent with the type of error that it introduces: heteroscedasticity tends to reduce coefficient p values, but does not bias the coefficients, and coefficient p values have been used primarily for inter-coefficient comparison rather than as absolute measures. P values of correlation tests used in the model evaluation are unaffected.

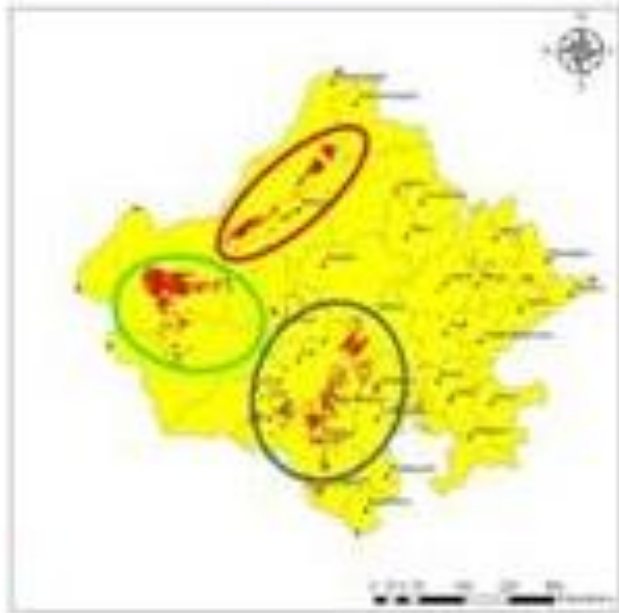


Figure 23: Rajasthan suitability map; most suitable area circled green. Original image quality. (Khan and Rathi, 2014)

In Karnataka, z-scores are lower than in Rajasthan, but more correlations are significant, including for area-adjusted values ([Table 11](#)). This suggests models are more accurately differentiating levels of suitability. MWR shows intelligible patterns of coefficient variation at sub-regional level. Where these patterns corroborate the extended hypothesis they could be used to enhance predictiveness, removing contextual effects through better model specification (Fotheringham *et al.*, 2002: 24). Log transforming increases the statistical validity of the models, while HSTR allows for regionally tailored regressions, further improving the goodness of fit ([Table 11](#); [Figure 17](#), [Figure 18](#)). Improved outcomes in Karnataka suggest that the methods used are more suited to areas with widely distributed SEG, possibly reflected in the lower heteroscedasticity of residuals (Appendix [9.14](#)). The apparent paucity of underexploited high-suitability areas suggests Karnataka's sites have been well-chosen. Though this reasoning may appear circular, as site data determined the suitability indices, other outcomes are possible: in Rajasthan, criteria inferred from current sites consistently highlight low-SEG western Jaisalmer as superior to the present 90 % polygon.

While intercorrelation introduces confounding elements, these are measured correlations between independent variables, not merely artefacts introduced by the method. Awareness of them is important for the construction of appropriate suitability indices.

6 Conclusions

Regression analysis of existing SEG has been used to infer SEG siting criteria in Karnataka and Rajasthan. While inferences about motivation are necessarily uncertain, criterion weights and suitability maps produced broadly cohere with academic consensus and ground truth, raising confidence in the validity of the intermediate steps. Maps generated suggest that Jaisalmer in Rajasthan and the areas around Niragunda, Patapalya and the far south of Karnataka are being underexploited, though findings for Rajasthan are of lower confidence.

Regarding the basic hypothesis, coefficient signs of Population, Irradiance, Powerlines, Slope and AODb generally agree in both states with the influence ascribed to these variables in the literature, while that of Temperature does not, suggesting a topic for further research. No conclusion was reached for Roads. Most variables are more coherent where they are more limiting, corroborating the extended hypothesis. This finding may be of use in the construction of suitability indices in other regions, though its generality is limited by intercorrelation of variables.

In Karnataka, spatial patterns of criterion importance have been demonstrated through regional and MWR analysis. These include approximately linear latitudinal variation and distinct north/south regimes, and highlight the importance of locally tailored suitability indices. In the cases of Irradiance and Population, much stronger correlation of β^* with latitude than with variable magnitude underlines the importance of spatial analysis. Future research could test the generality of the extended hypothesis and explore the causes of these spatial relationships.

HSTR has been used to make suitability indices synthesising regression outputs from multiple regions, and maps based on these combined indices have been shown to achieve better correlation with SEG data than those based on global regression. This shows that regional specificity is not incompatible with broad coverage. The requirement of a rich, spatially distributed dataset for HSTR and MWR is a methodological limitation.

The coherence of the results suggests that regression analysis offers meaningful insight into local siting criteria. The key finding of this study is that data-driven methods can provide an effective empirical complement to expertise-based suitability assessment techniques.

7 Acknowledgements

My sincere thanks go to Dr Amii Harwood, who supervised this project, for her consistent and invaluable guidance. Professor Andrew Lovett provided key clarification on a number of questions around appropriate spatial statistics, for which I am also very grateful. The Rajasthan Solar Association supplied data without which the project would have been much weaker, and I thank them for it. Errors are my own.

8 References

- Adhikari, D., 2020. Policy review and analysis promoting solar-powered irrigation in India. *Half Yearly Technical Journal of Indian Geographical Committee of IWRA*, **9**(1), 8-15.
- Ajmer Vidyut Vitran Nigam Limited, 2019. *11 kV Feeder Wise Cumulative Energy Audit Report of Ajmer Discom from Apr-18 to Dec-18*. Available at https://energy.rajasthan.gov.in/content/dam/raj/energy/avvn/pdf/EnergyAuditReport-2017/EA_REPORT_WTA_APRIL-18_to_DEC-18_ATC.pdf (last accessed 01/04/2021)
- Akkas, O.P., Erten, M.Y., Cam, E. and Inanc, N., 2017. Optimal Site Selection for a Solar Power Plant in the Central Anatolian Region of Turkey. *International Journal of Photoenergy*. doi:10.1155/2017/7452715.
- Al Garni, H.Z. and Awasthi, A., 2017. A fuzzy AHP and GIS-based approach to prioritize utility-scale solar PV sites in Saudi Arabia. *2017 IEEE International Conference on Systems, Man, and Cybernetics*, Banff, Alberta, Canada, 1244-1249. doi: 10.1109/SMC.2017.8122783
- Albadwawi, O., John, J., Elhassan, Y., Albanna, F., Almamari, F., Alqassim, A., and Alnuaimi, A., 2019. Quantification of spectral losses of natural soiling and detailed microstructural analysis of dust collected from different locations in Dubai, UAE. *IEEE 46th Photovoltaic Specialists Conference*, Chicago, IL, USA, 3115-3118. doi: 10.1109/PVSC40753.2019.8980937.
- Alharthi, Y.Z., Siddiki, M.K., and Chaudhry, G.M., 2018. Resource assessment and techno-economic analysis of a grid-connected solar PV-wind hybrid system for different locations in Saudi Arabia. *Sustainability*, **10**(10), 3690. doi:org/10.3390/su10103690
- Allen, M., Dube, O.P., Solecki, W., Aragón-Durand, F., Cramer, W., Humphreys, S., Kainuma, M., Kala, J., Mahowald, N., Mulugetta, Y., Perez, R., Wairiu, M., Zickfeld, K., 2018. Framing and Context. In: Masson-Delmotte, V., Zhai, P., Pörtner, H.O., Roberts, D., Skea, J., Shukla, P.R., Pirani, A., Chen, Y., Connors, S., Gomis, M., Lonnoy, E., Matthews, J.B.R., Moufouma-Okia, W., Péan, C., Pidcock, R., Reay, N., Tignor, M., Waterfield, T. and Zhou, X. (Eds.), *Global warming of 1.5°C. An IPCC special report on the impacts of global warming of 1.5°C above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change, sustainable development, and efforts to eradicate poverty*. In press.
- Aly, A., Jensen, S. and Pedersen, A., 2017. Solar power potential of Tanzania: identifying CSP and PV hot spots through a GIS multicriteria decision making analysis. *Renewable Energy*, **113**. doi: 10.1016/j.renene.2017.05.077
- Ammar, H., Benbaha, N. and Boukebbous, S.E., 2019. Siting analysis of PV water pumping system using GIS-based fuzzy analytic hierarchy process. *Proceedings of the 4th International Conference on Power Electronics and their Applications*, 1-5. doi: 10.1109/ICPEA1.2019.8911171
- Arthur, W., and Arrow, K., 1994. *Increasing Returns and Path Dependence in the Economy*. University of Michigan Press, Ann Arbor. doi: 10.3998/mpub.10029
- Arvizu, D., Balaya, P., Cabeza, L., Hollands, T., Jäger-Waldau, A., Kondo, M., Konseibo, C., Meleshko, V., Stein, W., Tamaura, Y., Xu, H., and Zilles, R., 2011. Direct Solar Energy. In: Edenhofer, O., Pichs-Madruga, R., Sokona, Y., Seyboth, K., Matschoss, P., Kadner, S., Zwickel, T., Eickemeier, P., Hansen, G., Schlömer, S., von Stechow, C. (Eds.), *IPCC Special Report on Renewable Energy Sources and Climate Change Mitigation*. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 333-400

- Azadeh, A., Ghaderi, S.F. and Maghsoudi, A., 2006. Location Optimization of Solar Plants by an Integrated Multivariable DEA-PCA Method. *2006 IEEE International Conference on Industrial Technology*, Mumbai, India, 1867-1872. doi: 10.1109/ICIT.2006.372571.
- Azevêdo, V.W.B., Tiba, C., Candeias, L.B., 2017. Location study of solar thermal power plant in the state of Pernambuco using geoprocessing technologies and multiple-criteria analysis. *Energies*, **10**(7). doi: 10.3390/en10071042
- Balta-Ozkan, N., Yildirim, J. and Connor, P., 2015. Regional distribution of photovoltaic deployment in the UK and its determinants: A spatial econometric approach. *Energy Economics*, **51**, 417-429. doi: 10.1016/j.eneco.2015.08.003
- Balta-Ozkan, N., Yildirim, J., Connor, P., Truckell, I. and Hart, P., 2021. Energy transition at local level: Analyzing the role of peer effects and socio-economic factors on UK solar photovoltaic deployment. *Energy Policy*, **148**(B). doi: 10.1016/j.enpol.2020.112004
- Bensch, G., Kluge, J. and Peters, J., 2011. *Impacts of Rural Electrification in Rwanda*. Ruhr Economic Papers no. 284. ISBN 978-3-86788-330-6, Rheinisch-Westfälisches Institut für Wirtschaftsforschung, Essen
- Bridge to India, 2020a. *India Solar Map 2019 December*. Available at <https://bridgetoindia.com/backend/wp-content/uploads/2020/05/BRIDGE-TO-INDIA-India-Solar-Map-December-2019.pdf> (last accessed 14/04/2021)
- Bridge to India, 2020b. *India Solar Rooftop Map 2019 December*. Available at https://bridgetoindia.com/backend/wp-content/uploads/2020/05/BRIDGE-TO-INDIA-India-solar-rooftop-map_Dec-2019.pdf (last accessed 14/04/2021)
- Briguglio, M. and Formosa, G., 2017. When households go solar: Determinants of uptake of a photovoltaic scheme and policy insights. *Energy Policy*, **108**, 154-162
- Brownson, J.R.S., 2014. *Solar Energy Conversion Systems*. Academic Press. ISBN 9780123970213
- Central Electricity Authority, 2020. *Report on optimal generation capacity mix for 2029-30*. Government of India Ministry of Power. Available at https://cea.nic.in/old/reports/others/planning/irp/Optimal_mix_report_2029-30_FINAL.pdf (last accessed 25/02/2021)
- Chan, Y., 2011. *Location Theory and Decision Analysis* (2nd Edition). Springer-Verlag, Berlin/Heidelberg
- Charabi, Y. and Gastli, A., 2012. Spatio-temporal assessment of dust risk maps for solar energy systems using proxy data. *Renewable Energy*, **44**, 23-31.
- Charabi, Y. and Gastli, A., 2013. Integration of temperature and dust effects in siting large PV power plant in hot arid area. *Renewable Energy*, **57**, 635-644
- Chavan, G.B., Chaudhary, V.R. and Badhe, S.S., 2019. Site Suitability Analysis of Solar Farm using AHP Technique. *International Journal of Advance Engineering and Research Development*, **6**(11)
- Chen, C.-R., Huang, C.-C. and Tsuei, H.-J., 2014. A hybrid MCDM model for improving GIS-based solar farms site selection. *International Journal of Photoenergy*, 1-9. doi: 10.1155/2014/925370
- Colak, H.E., Memisoglu, T. and Gercek, Y., 2020. Optimal site selection for solar photovoltaic (PV) power plants using GIS and AHP: A case study of Malatya Province, Turkey. *Renewable Energy*, **149**, 565-576. doi: 10.1016/j.renene.2019.12.078

- Copiello, S. and Grillenzoni, C., 2017. Solar photovoltaic energy and its spatial dependence. *Energy Procedia*, **141**, 86-90. doi: 10.1016/j.egypro.2017.11.017
- Corcelli, F., Ripa, M. and Ulgiati, S., 2017. End-of-life treatment of crystalline silicon photovoltaic panels. An emergy-based case study. *Journal of Cleaner Production*, **161**, 1129-1142
- Database of Global Administrative Areas (GADM), 2012. *GADM database of Global Administrative Areas, version 2.5*. Available at <https://gadm.org/data.html> or via <https://diva-gis.org> (last accessed 16/04/2021)
- Defense Mapping Agency (DMA), 1992. Digital Chart of the World. Defense Mapping Agency, Fairfax, Virginia. Available via <https://diva-gis.org> (last accessed 16/04/2021)
- Dharshing, S., 2017. Household dynamics of technology adoption: A spatial econometric analysis of residential solar photovoltaic (PV) systems in Germany. *Energy Research & Social Science*, **23**, 113-124
- Directorate Of Census Operations Andhra Pradesh, 2011. *District Census Handbook Anantapur: Village and Town Directory*. Directorate Of Census Operations, Andhra Pradesh. Available at <https://censusindia.gov.in/2011census/dchb/ApBookA.html> (last accessed 11/04/2021)
- Doljak, D. and Stanojević, G., 2017. Evaluation of natural conditions for site selection of ground-mounted photovoltaic power plants in Serbia. *Energy*, **127**, 291-300
- Effat, H.A., 2013. Selection of potential sites for solar energy farms in Ismailia Governorate, Egypt using SRTM and multicriteria analysis. *International Journal of Advanced Remote Sensing and GIS*, **2**(1), 205-220
- Energy Department, Government of Rajasthan, 2019. *Rajasthan Solar Energy Policy, 2019*. Available at https://06668763-51b3-4994-951b-ec1dae2cc6c3.filesusr.com/ugd/af393a_b9d2072018f94576b19e2bee22481463.pdf (last accessed 15/04/2021)
- Energy Sector Management Assistance Program (ESMAP), 2019. *Global Solar Atlas 2.0 Technical Report*. World Bank, Washington, DC. Available at <http://documents1.worldbank.org/curated/en/529431592893043403/pdf/Global-Solar-Atlas-2-0-Technical-Report.pdf> (last accessed 16/04/2021)
- Environmental Systems Research Institute, 2012. *ArcGIS 10.1 Projected Coordinate System Tables*. Available at https://desktop.arcgis.com/en/arcmap/10.3/guide-books/map-projections/pdf/projected_coordinate_systems.pdf
- Environmental Systems Research Institute, 2018. *ArcGIS Desktop: Release 10.6.1*
- Flemming, P.G.K., 2020. A GIS investigation of photovoltaic generation siting in Andhra Pradesh. *ENV-5028B: GIS Skills for Project Work*, University of East Anglia. Unpublished essay.
- Forster, P., Ramaswamy, V., Artaxo, P., Bernsten, T., Betts, R., Fahey, D.W., Haywood, J., Lean, J., Lowe, D.C., Myhre, G., Nganga, J., Prinn, R., Raga, G., Schulz, M. and Van Dorland, R., 2007. Changes in atmospheric constituents and in radiative forcing. In: Solomon, S., Qin, D., Manning, M., Chen, Z., Marquis, M., Averyt, K.B., Tignor, M. and Miller, H.L. (Eds.), *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 129-234
- Fotheringham, S.A., Brunsdon C. and Charlton, M., 2002. *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. John Wiley & Sons Ltd, Chichester

- Freire, M.C., Pica, C.Q., Martins, M.A.I.; Silva, A.L.D., 2015. Analysis of relevant technical and economic aspects to support the choice of feasible locations for photovoltaic power plants in Brazil. *2015 IEEE PES Innovative Smart Grid Technologies Latin America (ISGT LATAM)*, Montevideo, Uruguay. 258-263. doi: 10.1109/ISGT-LA.2015.7381164
- Friedlingstein, P., O'Sullivan, M., Jones, M.W., Andrew, R.M., Hauck, J., Olsen, A., Peters, G.P., Peters, W., Pongratz, J., Sitch, S., Le Quéré, C., Canadell, J.G., Ciais, P., Jackson, R.B., Alin, S., Aragão, L.E.O.C., Arneeth, A., Arora, V., Bates, N.R. and Becker, M., 2020. Global Carbon Budget 2020. *Earth System Science Data*, **12**(4), 3269-3340
- Gastli, A., and Charabi, Y., 2010. Siting of large PV farms in Al-Batinah region of Oman. *2010 IEEE International Energy Conference Energy Conference and Exhibition (EnergyCon)*, 548-552
- Ghose, D., Naskar, S., Shabbiruddin, Sadeghzadeh, M., Assad, M.E.H. and Nabipour, N., 2020. Siting high solar potential areas using Q-GIS in West Bengal, India. *Sustainable Energy Technologies and Assessments*, **42**
- Giamalaki, M. and Tsoutsos, T., 2019. Sustainable siting of solar power installations in Mediterranean using a GIS/AHP approach. *Renewable Energy*, **141**, 64-75
- Ginley, D., Ager, J., Agrawal, R., Alam, M.A., Arora, B.M., Avasthi, S., Basak, D., Bhargava, P., Biswas, P., Bora, B., Braunecker, W.A., Buonassisi, T., Dhage, S., Dhere, N., Garner, S., Hu, X., Jhunjhunwala, A., Kabra, D., Kavaipatti, B., Kazmerski, L. *et al.*, 2020. Sustainable Photovoltaics. In: Ginley, D. and Chattopadhyay, K. (Eds.), *Solar Energy Research Institute for India and the United States (SERIUS)*. Lecture Notes in Energy 39. Springer Nature Switzerland, 25-85
- Government of Rajasthan Panchayati Raj, 2019. *Allotment for the year 2018-2019*. Available at <http://rajpanchayat.rajasthan.gov.in/en-us/announcementcirculars/circulars/accounts/allotmentfortheyear201819.aspx> (last accessed 10/04/2021)
- Greenland, S., Morgenstern, H., Maclure, M., Schlesselman, J.J., and Poole, C., 1991. Standardized Regression Coefficients: A Further Critique and Review of Some Alternatives. *Epidemiology*, **2**(5), 387-392
- Gujarati, D.N., 2019. The linear regression model. In: *Linear Regression: A Mathematical Introduction*. SAGE Publications, Thousand Oaks, 1-21. doi: 10.4135/9781071802571
- Hertwich, E.G., Gibon, T., Bouman, E.A., Arvesen, A., Suh, S., Bergesen, J.D., Heath, G.A., Ramirez, A., Vega, M.I. and Shi, L., 2015. Integrated life-cycle assessment of electricity-supply scenarios confirms global environmental benefit of low-carbon technologies. *Proceedings of the National Academy of Sciences of the United States of America*, **112**(20), 6277-6282
- Indian Village Boundaries Project, 2017. *indian_village_boundaries/karnataka*. Data{Meet}. Available at https://projects.datameet.org/indian_village_boundaries/ka/ under the [Open Database License](#) (last accessed 16/04/2021)
- Indian Village Boundaries Project, 2018. *indian_village_boundaries/rajasthan*. Data{Meet}. Available at https://github.com/tejesh0/indian_village_boundaries/tree/rajasthan/rj under the [Open Database License](#) (last accessed 16/04/2021)
- International Energy Agency, 2020a. *World Energy Outlook 2020*. OECD Publishing, Paris. doi: 10.1787/557a761b-en.
- International Energy Agency, 2020b. *India 2020: Energy Policy Review*. OECD Publishing, Paris. doi: 10.1787/9faa9816-en

International Energy Agency, 2020c. *Electricity Access Database*. International Energy Agency, Paris. Available at <https://iea.blob.core.windows.net/assets/93fd1a56-5c8f-4209-ba6e-7f6ff9fffb19/WEO2020-Electricityaccessdatabase.xlsx> (last accessed 16/04/2021)

International Energy Agency and World Bank, 2018. *Sustainable Energy for All (SE4ALL) Database*. Available at <http://api.worldbank.org/v2/en/indicator/EG.ELC.ACCS.ZS?downloadformat=excel> (last accessed 16/04/2021)

International Institute for Sustainable Development, 2017. *India's Energy Transition: Mapping Subsidies to Fossil Fuels and Clean Energy in India*. International Institute for Sustainable Development, Winnipeg

Izeiroski, S., Kotevska, E., Panovski, S. and Nedelkovski, I., 2016. Contemporary GIS-based methodological approach for assessment of optimal locations for exploitation of solar-energy potentials. *Micro, Macro & Mezzo Geoinformation*, **7**, 23-38

Jain, A.K., Briegleb, B.P., Minschwaner, K. and Wuebbles, D.J., 2000. Radiative forcings and global warming potentials of 39 greenhouse gases. *Journal of Geophysical Research*, **105** (D16), 20773-20790

Jan, I., Ullah, W. and, Ashfaq, M., 2020. Social acceptability of solar photovoltaic system in Pakistan: Key determinants and policy implications. *Journal of Cleaner Production*, **274**

Jawaharlal Nehru National Solar Mission, 2009. *Towards Building Solar India*. Available at [https://www.seci.co.in/upload/static/files/mission_document_JNNSM\(1\).pdf](https://www.seci.co.in/upload/static/files/mission_document_JNNSM(1).pdf) (last accessed 15/04/2021)

Jessen, W., Wilbert, S., Gueymard, C.A., Polo, J., Ramírez, L., Bian, Z., Driesse, A., Habte, A., Marzo, A., Armstrong, P.R., and Vignola, F., 2018. Proposal and evaluation of subordinate standard solar irradiance spectra for applications in solar energy systems. *Solar Energy*, **168**, 30-43

Jodhpur Vidyut Vitran Nigam Limited, 2020. *AT&C Loss Report (With Arrear) - HTFeeder Wise 01-04-2018 to 31-03-2019*. Available at <https://energy.rajasthan.gov.in/content/raj/energy-department/jodhpur-vidyut-vitran-nigam-limited/en/jdvvn-corner/feeder-wise-at-c-losses.html#> (last accessed 01/04/2021)

Jodhpur Vidyut Vitran Nigam Limited, 2020. *AT&C Loss Report (With Arrear) - HTFeeder Wise 01-04-2017 to 31-03-2018*. Available at <https://energy.rajasthan.gov.in/content/raj/energy-department/jodhpur-vidyut-vitran-nigam-limited/en/jdvvn-corner/feeder-wise-at-c-losses.html#> (last accessed 01/04/2021)

Jun, D., Tian-tian, F., Yi-sheng, Y. and Yu, M., 2014. Macro-site selection of wind/solar hybrid power station based on ELECTRE-II. *Renewable and Sustainable Energy Reviews*, **35**, 194-204. doi: 10.1016/j.rser.2014.04.005

Karnataka Electricity Regulatory Commission, 2019a. *Determination of Tariff in Respect of Solar Power Projects (Including Solar Rooftop Photovoltaic Projects) for FY20*. Available at [https://karunadu.karnataka.gov.in/kerc/Documents/Determination%20of%20tariff%20in%20respect%20of%20Solar%20Power%20Projects%20\(including%20Solar%20Rooftop%20Photovoltaic%20Projects%20for%20FY20.pdf](https://karunadu.karnataka.gov.in/kerc/Documents/Determination%20of%20tariff%20in%20respect%20of%20Solar%20Power%20Projects%20(including%20Solar%20Rooftop%20Photovoltaic%20Projects%20for%20FY20.pdf) (last accessed 08/08/2020)

Karnataka Electricity Regulatory Commission, 2019b. *Decision on Various Models and Guidelines for Solar Rooftop Photovoltaic Plants allowed to be installed on rooftops of the consumers' buildings*. Available at <https://karunadu.karnataka.gov.in/kerc/Documents/Decision%20on%20various%20models%20and%20guidelines%20for%20solar%20rooftop%20photovoltaic%20plants%20allowed%20to%20be%20installed%20on%20rooftops%20of%20the%20consumers'%20buildings.pdf>

[20Guidelines%20for%20SRTPV%20allowed%20to%20be%20installed%20on%20rooftops%20of%20consumer%20buildings.pdf](#) (last accessed 16/04/2021)

Karnataka Power Transmission Corporation Limited, 2020. *ESCOM Wise Load*. Available at <http://kptclsltd.com/escom.aspx> (last accessed 31/10/2020)

Karnataka Renewable Energy Development Limited, 2019. *Details of Solar Power Project Allotted/Commissioned under various category in Karnataka*. Available at http://kredlinfo.in/solargrid/sglist/Solar_alotlist.xlsx (last accessed 16/04/2021)

Kengpol, A., Rontlaong, P. and Tuominen, M., 2013. A decision support system for selection of solar power plant locations by applying fuzzy AHP and TOPSIS: An empirical study. *Journal of Software Engineering and Applications*, **6**(9), 470-481. doi: 10.4236/jsea.2013.69057

Khan, G. and Rathi, S., 2014. Optimal site selection for solar PV power plant in an Indian state using geographical information system (GIS). *International Journal of Emerging Engineering Research and Technology*, **2**(7), 260-266

Kim, S, Lee, Y. and Moon, H.-R., 2018. Siting criteria and feasibility analysis for PV power generation projects using road facilities. *Renewable and Sustainable Energy Reviews*, **81**, 3061-3069

Kline, R.B., 2016. *Principles and Practice of Structural Equation Modeling* (4th Edition). The Guilford Press, New York and London

Koc, A., Turk, S. and Şahin, G., 2019. Multi-criteria of wind-solar site selection problem using a GIS-AHP-based approach with an application in Igdir Province/Turkey. *Environmental Science and Pollution Research*, **26**, 32298-32310

Kumar, A., Prakash, O. and Dubec, A., 2017. A review on progress of concentrated solar power in India. *Renewable and Sustainable Energy Reviews*, **79**, 304-307

Laarabi, B., El Baqqal, Y., Dahrouch, A. and Barhdadi, A., 2020. Deep analysis of soiling effect on glass transmittance of PV modules in seven sites in Morocco. *Energy*, **213**

Lee, M. and Hong, T., 2019. Hybrid agent-based modeling of rooftop solar photovoltaic adoption by integrating the geographic information system and data mining technique. *Energy Conversion and Management*, **183**, 266-279

Levy, R.C., Hsu, C., et al., 2015. *MODIS Atmosphere L2 Aerosol Product*. NASA MODIS Adaptive Processing System, Goddard Space Flight Center, USA. doi: 10.5067/MODIS/MOD04_L2.061

Lin, A., Ming Lu, M. and Li, C., 2017. On spatial distribution and determinants of urban photovoltaic utilization in China. *Energy Procedia*, **134**, 470-479

Lloyd, C.T., Sorichetta, A., and Tatem, A.J., 2017. Data Descriptor: High resolution global gridded data for use in population studies. *Scientific Data*, **4**. doi: 10.1038/sdata.2017.1

Lobo, S. and Bhavani, R.G., 2018. Performance and comparative analysis of a medium sized, grid connected photovoltaic system for some locations in the UAE. *2018 IEEE Innovative Smart Grid Technologies - Asia (ISGT Asia)*, 1215-1220. doi: 10.1109/ISGT-Asia.2018.8467951

Mardani, A. Jusoh, A., Nor, K.M.D., Khalifah, Z., Zakwan, N. and Valipour, A., 2015a. Multiple criteria decision-making techniques and their applications – a review of the literature from 2000 to 2014. *Economic Research-Ekonomska Istraživanja*, **28**(1), 516-571, doi: 10.1080/1331677X.2015.1075139

- Mardani, A., Jusoh, A., Zavadskas, E.K., Cavallaro, F. and Khalifah, Z., 2015b. Sustainable and renewable energy: an overview of the application of multiple criteria decision making techniques and approaches. *Sustainability*, **7**, 13947-13984. doi: 10.3390/su71013947
- MATLAB, 2017. *version 7.10.0 (R2017b)*. The MathWorks Inc, Massachusetts
- Mensour, O.N., El Ghazzani, B., Hlimi, B. and Ihlal, A., 2019. A geographical information system-based multi-criteria method for the evaluation of solar farms locations: A case study in Souss-Massa area, southern Morocco. *Energy*, **182**, 900-919
- Merrouni, A.A., Mezrhah, A. and Mezrhah, A., 2016. PV sites suitability analysis in the Eastern region of Morocco. *Sustainable Energy Technologies and Assessments*, **18**, 6-15. doi:10.1016/j.seta.2016.09.006
- Microsoft Research, 2020. *Fuzzy Lookup Add-in for Excel*. Available at <https://www.microsoft.com/en-gb/download/details.aspx?id=15011> (last accessed 18/04/2021)
- Mierzwiak, M. and Calka, B., 2017. Multi-Criteria Analysis for Solar Farm Location Suitability. *Reports on Geodesy and Geoinformatics*, **104**(1), 20-32. doi: 10.1515/rgg-2017-0012
- NASA JPL, 2013. *NASA Shuttle Radar Topography Mission Global 1 arc second SRTMGL1v003*. NASA EOSDIS Land Processes DAAC. doi: 10.5067/MEaSUREs/SRTM/SRTMGL1.003. Available at <https://lpdaac.usgs.gov/products/srtmgl1v003/> (last accessed 16/04/2021)
- Newman, T.B. and Browner, W.S., 1991. In defense of standardised regression coefficients. *Epidemiology*, **2**(5), 383-386
- Nieminen, P., Lehtiniemi, H., Vähäkangas, K., Huusko, A. and Rautio, A. Standardised regression coefficient as an effect size index in summarising findings in epidemiological studies. *Epidemiology Biostatistics and Public Health*, **10**(4)
- Nikšić, H., 2015. GNU Wget: *The Non-interactive Download Utility*. Samurai Media
- OpenStreetMap contributors, 2020. *Highways data*. Available at <https://mapcruzin.com/free-india-country-city-place-gis-shapefiles.htm> (last accessed 15/03/2021)
- OpenStreetMap contributors, 2020. *Powerline data*. Retrieved via <https://data.nextgis.com> (last accessed 21/01/2021)
- Palmer, D., Ralph Gottschalg, R. and Betts, T., 2020. The future scope of large-scale solar in the UK: Site suitability and target analysis. *Renewable Energy*, **133**, 1136-1146
- Pek, J., Wong, O. and Wong, A.C.M., 2018. How to address non-normality: A taxonomy of approaches, reviewed, and illustrated. *Frontiers in Psychology*, **9**
- Peter, R., Dickie, L. and Peter, V.M., 2006. Adoption of photovoltaic power supply systems: A study of key determinants in India. *Renewable Energy*, **31**, 2272-2283
- Rajasthan Rajya Vidyut Prasaran Nigam Limited (RRVPL), 2012. *India: Rajasthan Renewable Energy Transmission Investment Program (Tranche -1): Initial Environmental Examination*. Government of Rajasthan. Available at <https://www.adb.org/sites/default/files/project-document/73171/45224-002-ind-ieee-01-draft.pdf> (last accessed 18/02/2021)
- Rediske, G., Siluk, J.C.M., Rigo, P.D., Michels, L., Rosa, C.B. and Cugler, G., 2020. Multi-criteria decision-making model for assessment of large photovoltaic farms in Brazil. *Energy*, **197**
- Roy, P.S., Meiyappan, P., Joshi, P.K., Kale, M.P., Srivastav, V.K., Srivasatava, S.K., Behera, M.D., Roy, A., Sharma, Y., Ramachandran, R.M., Bhavani, P., Jain, A.K., and Krishnamurthy, Y.V.N., 2016. *Decadal*

- Land Use and Land Cover Classifications across India, 1985, 1995, 2005*. ORNL DAAC, Oak Ridge, Tennessee, USA. doi: 10.3334/ORNLDAAC/1336
- Ruiz, H.S., Sunarso, A., Ibrahim-Bathis, K., Murti, S.A. and Budiarto, I., 2020. GIS-AHP multi criteria decision analysis for the optimal location of solar energy plants at Indonesia. *Energy Reports*, **6**, 3249-3263
- Saaty, T.L., 1994. How to make a decision: the analytic hierarchy process. *Interfaces*, **24**(6), 19-43. doi: 10.1287/inte.24.6.19
- Samanlioglu, F. and Ayağ, Z., 2017. A fuzzy AHP-PROMETHEE II approach for evaluation of solar power plant location alternatives in Turkey. *Journal of Intelligent & Fuzzy Systems*, **33**, 859-871. doi: 10.3233/JIFS-162122
- Samu, R., Poyrazoglu, G. and Fahrioglu, M., 2019. The potential and economic analysis of grid-connected solar PV power in Kenya. *1st Global Power, Energy and Communication Conference (GPECOM), Nevsehir, Turkey*. 298-301. doi: 10.1109/GPECOM.2019.8778467
- Sánchez-Lozano, J. M., Teruel-Solano, J., Soto-Elvira, P.L., and Socorro García-Cascales, M., 2013. Geographical Information Systems (GIS) and Multi-Criteria Decision Making (MCDM) methods for the evaluation of solar farms locations: Case study in south-eastern Spain. *Renewable and Sustainable Energy Reviews*, **24**, 544-556. doi: 10.1016/j.rser.2013.03.019
- Saraswat, S.K. Digalwar, A.K., Yadav, S.S. and Kumar, G., 2021. MCDM and GIS based modelling technique for assessment of solar and wind farm locations in India. *Renewable Energy*, **169**, 865-884
- Schroeder, L., Sjoquist, D. and Stephan, P., 2017. Multiple linear regression. In: Schroeder, L., Sjoquist, D. and Stephan, P. (Eds.), *Understanding Regression Analysis* (Second Edition). SAGE Publications, Inc, 21-30
- Schunder, T., Yin, D., Bagchi-Sen, S. and Rajan, K., 2020. A spatial analysis of the development potential of rooftop and community solar energy. *Remote Sensing Applications: Society and Environment*, **19**
- Seppaelae, U., 1997. *An Evolutionary Model for Spatial Location of Economic Facilities*. IIASA Interim Report.
- Shrimali, G. and Rohra, S., 2012. India's solar mission: A review. *Renewable and Sustainable Energy Reviews*, **16**, 6317-6332
- Sindhua, S., Nehra, V., Luthra, S., 2017. Investigation of feasibility study of solar farms deployment using hybrid AHP-TOPSIS analysis: Case study of India. *Renewable and Sustainable Energy Reviews*, **73**, 496-511
- Silverman, B.W., 1986. *Density estimation for statistics and data analysis*. Chapman and Hall, London
- Skoplaki, E. and Palyvos, J.A., 2009. On the temperature dependence of photovoltaic module electrical performance: A review of efficiency/power correlations. *Solar Energy*, **83**, 614-624
- SolarGis, 2019. *Global Solar Atlas 2.0 (India)* (Operated by SolarGis s.r.o on behalf of the World Bank Group, utilizing SolarGis data, with funding provided by the Energy Sector Management Assistance Program (ESMAP)). Available at <https://solargis.com/maps-and-gis-data/download/india> (last accessed 16/04/2021)
- Subramanya, U.R., 2012. District Domestic Product of Karnataka. In: Subramanya, U. R., Jayaram, D., and Muniraju, V., 2013. *Consolidated Articles*, 193-217

- Suh, J. and Brownson, J.R.S., 2016. Solar farm suitability using geographic information system fuzzy sets and analytic hierarchy processes: Case study of Ulleung Island, Korea. *Energies*, **9**. doi: 10.3390/en9080648
- Surve, B.C., 2014. Web based spatio-temporal data modeling and application development for solar power plant siting using GIS for India. *Proceedings of the 2014 Conference on IT in Business, Industry and Government: An International Conference by CSI on Big Data, CSIBIG*. Institute of Electrical and Electronics Engineers Inc. doi: 10.1109/CSIBIG.2014.7056968
- Tahri, M., Mustapha Hakdaoui, M. and Maanan, M., 2015. The evaluation of solar farm locations applying geographic information system and multi-criteria decision-making methods: Case study in southern Morocco. *Renewable and Sustainable Energy Reviews*, **51**, 1354-1362
- Tavana, M., Santos Arteaga, F.J., Mohammadi, S. and Alimohammadi, M., 2017. A fuzzy multi-criteria spatial decision support system for solar farm location planning. *Energy Strategy Reviews*, **18**, 93-105
- Teräsvirta, T., 1994. Specification, estimation, and evaluation of smooth transition autoregressive models. *Journal of the American Statistical Association*, **89**(425), 208-218
- Tercan, E., Saracoglu, B.O., Bilgilioglu, S.S., Eymen, A. and Tapkin, S., 2020. Geographic information system-based investment system for photovoltaic power plants location analysis in Turkey. *Environmental Monitoring & Assessment*, **192**(5), 1-26. doi: 10.1007/s10661-020-08267-5
- Tiba, C. and Beltrão, R.E. de A., 2012. Siting PV plant focusing on the effect of local climate variables on electric energy production - Case study for Araripina and Recife. *Renewable Energy*, **48**, 309-317
- UN General Assembly, 2015. *Transforming our World: the 2030 Agenda for Sustainable Development*. A/RES/70/1. Available at https://www.un.org/ga/search/view_doc.asp?symbol=A/RES/70/1&Lang=E (last accessed 16/04/2021)
- Uyan, M., 2017. Optimal site selection for solar power plants using multi-criteria evaluation: A case study from the Ayranci region in Karaman, Turkey. *Clean Technologies and Environmental Policy*, **19**(9), 2231-2244. doi: 10.1007/s10098-017-1405-2
- Van Rossum, G. and Drake Jr, F.L., 1995. *Python reference manual*. Centrum voor Wiskunde en Informatica, Amsterdam
- Waewsak, J., Chancham, C., Ali, S., Natee, W., Kongruang, C. and Gagnon, Y., 2020. Assessment of hybrid, firm renewable energy-based power plants: Application in the southernmost region of Thailand. *Renewable and Sustainable Energy Reviews*, **130**. doi: 10.1016/j.rser.2020.109953
- Wilcox, R.R., 2017. *Introduction to Robust Estimation and Hypothesis Testing* (4th Edition). Academic Press, London and Oxford
- Wildlife Institute of India, undated. *Wildlife Institute of India Geographic Data*. Available at http://210.212.84.122:80/erdas-apollo/vector/WII_GEOGRAPHIC_DATA (last accessed 13/11/2020)
- Wong, D., 2009. The Modifiable Areal Unit Problem (MAUP). In: Fotheringham, A.S. and Rogerson, P.A. (Eds.), *The SAGE Handbook of Spatial Analysis*. SAGE Publications, London, 105-124
- WorldPop and Center for International Earth Science Information Network (CIESIN), 2018. *Global High Resolution Population Denominators Project - Funded by The Bill and Melinda Gates Foundation*

(OPP1134076). doi: 10.5258/SOTON/WP00660. Columbia University. Creative Commons Attribution 4.0 International Licence. Available at <https://www.worldpop.org/> (last accessed 16/04/2021)

Wu, Y., Geng, S., Zhang, H. and Gao, M., 2014. Decision framework of solar thermal power plant site selection based on linguistic Choquet operator. *Applied Energy*, **136**, 303-311

Yenneti., K., 2016. The grid-connected solar energy in India: Structures and challenges. *Energy Strategy Reviews*, **11-12**, 41-51

Zoghi, M., Ehsani, A.H., Sadat, M., Amiri, M.J. and Karimi, S., 2017. Optimization solar site selection by fuzzy logic model and weighted linear combination method in arid and semi-arid region: A case study Isfahan-IRAN. *Renewable and Sustainable Energy Reviews*. **68**, 986-996. doi: 10.1016/j.rser.2015.07.014

9 Appendices

9.1	SEG determinant studies reviewed	64
9.2	SEG siting studies reviewed	65
9.3	Economic factors considered in studies reviewed	67
9.4	Data sources	68
9.5	Kernel Density.....	69
9.6	AODb data code.....	70
9.6.1	Data access wget	70
9.6.2	Data extraction code	70
9.6.3	AODb data collation by season	72
9.7	Karnataka electricity consumption scraper code	74
9.8	Substation geocoding	74
9.9	Normality of input datasets.....	75
9.10	Distance to roads map layer, Rajasthan	76
9.11	Matlab Multilinear Regression code.....	76
9.12	Matlab Moving Window Regression code	78
9.12.1	Driver code (front end).....	78
9.12.2	Regression code (back end)	82
9.12.3	Helper function.....	83
9.13	Regression Equations.....	84
9.14	Normality and homoscedasticity of residuals	85
9.15	MWR correlation strengths	85
9.16	Weight standardisation for comparison with previous studies	86

9.1 SEG determinant studies reviewed

Table 13: Determinant studies. See [References](#) for full bibliographic details.

Lead author	Year	Country/ state	Region	SEG scale
Balta-Ozkan	2015	UK	Europe	Residential
Balta-Ozkan	2021	UK	Europe	Residential
Best	2019	Australia	Australia	Residential
Bollinger	2012	California	USA	Residential
Briguglio	2017	Malta	Europe	Residential
Copiello	2017	Italy	Europe	Residential
Dharshing	2017	Germany	Europe	Residential
Jan	2020	Pakistan	ISC	Residential
Lee	2019	Korea	Asia	Residential
Lin	2017	China	Asia	Distributed
Peter	2006	India	ISC	Distributed
Qureshi	2017	Pakistan	ISC	Residential
Robinson	2015	Texas	USA	Residential
Schunder	2020	New York	USA	Residential, community
Sriwannawita	2015	Bangladesh	ISC	Residential

Return to [2.1 Determinant studies](#)

9.2 SEG siting studies reviewed

Table 14: Studies reviewed.

Author key: *: not in weight comparison plot. †: CSP only. **External consultation key:** A: Academic. E: Economists. G: Government. I: Industry. D: Discussion only; authors make decisions. ND: No Data given. Where two figures are given for “number contributing”, the first refers to the number of authors, the second to the number of experts consulted. **Criterion key:** ✓: evaluated. X: exclusion criterion. m: mentioned. p: proxy used. Other abbreviations: [Table 15](#).

Paper overview				Methodology				Criteria																		Upper slope limit, °	upper population distance	
								Climate					Terrain						Infrastructure			Social						
Lead Author	Year	Country/ state	Region	MCDAs	GIS	External consultation	Number contributing	Irradiance	Temperature	Soiling	Humidity	Wind	LULC	Water	Slope	Aspect	Elevation	Available Area	Hazards	Roads	Powerlines	Substations	Population	Economic	Policy	Demand		
Akkas*	2017	Turkey	ME	AHP; others			4	✓							✓												9	
Al Garni	2017	Saudi Arabia	ME	fAHP	✓		2	✓	✓				X		✓	✓				✓	✓		✓				5	
Alharthi*	2018	Saudi Arabia	ME				3	✓																				
Aly	2017	Tanzania	SSA	AHP	✓	lit. rev.	3	✓					X	X						✓	✓		✓			p		45
Ammar	2019	Algeria	NA/Med	fAHP	✓		3	✓					X	✓						✓			X					
Azadeh	2006	Iran	ME				3	✓																				
Azevêdo†	2017	Brazil	SAm	AHP	✓		3	✓					✓	✓	✓					✓	✓	m	✓				3	57.2
Charabi	2013	Oman	ME	FLOWA	✓		2	✓	✓	✓										✓								
Chavan	2019	Pune	ISC	AHP	✓		3	✓					✓	✓	✓	✓				✓	✓						82	
Chen	2014	China	Asia	dANP	✓	experts: I, A, G	15	✓	✓				✓		✓	✓		✓		✓	✓	✓	✓					
Colak	2020	Turkey	ME	AHP	✓	experts	ND	✓					X	✓		✓			✓	✓	✓	✓	✓					
Doljak	2017	Serbia	NEu	AHP	✓		2	✓	✓		✓		✓	X	✓	✓											32	
Effat	2013	Egypt	ME	AHP	✓	D	1, ND	✓					X	X		✓				✓	✓		✓					
Freire*	2015	Brazil	SAm				4		X	X		X	✓								X							
Gastli	2010	Oman	ME	AHP-FLOWA	✓		2	✓		X			X	X	X					✓	X		X				3	
Ghose	2020	West Bengal	ISC	AHP	✓	experts: I, A, G	3	✓	✓		✓		X	X	✓					✓	✓						6	
Giamalaki	2019	Crete	NA/Med	AHP	✓	energy groups	ND	✓					✓	✓	✓	✓	✓			✓	✓						16	
Izeiroski	2016	Macedonia	NEu		✓		4	✓							X	X		X									3	
Jun	2014	China	Asia	ELECTRE-II		experts?	4, ND	✓												✓	✓			✓	p			
Kengpol	2013	Thailand	Asia	fAHP-TOPSIS	✓	unclear	3, ND	✓	0				X	X	0	X	X		0	✓		✓	0	✓	0			

Paper overview				Methodology				Criteria																	Upper slope limit, °	upper population distance		
								Climate					Terrain						Infrastructure			Social						
Lead Author	Year	Country/ state	Region	MCDA	GIS	External consultation	Number contributing	Irradiance	Temperature	Soiling	Humidity	Wind	LULC	Water	Slope	Aspect	Elevation	Available Area	Hazards	Roads	Powerlines	Substations	Population	Economic	Policy	Demand		
Khan*	2014	Rajasthan	ISC		✓		2	✓		X			X		✓					✓	✓							
Kim	2018	Korea	Asia		✓		3						X	X	X			X		X	X			✓				6
Koc	2019	Turkey	ME	AHP	✓	unclear	3	✓	✓			✓	✓		✓	✓	✓				✓							3
Lobo*	2018	UAE	ME				2	✓	✓	✓																		
Merrouni	2018	Morocco	NA/Med	AHP	✓	D (I?)	3, ND	✓	0				X	✓	✓					✓	✓		✓					3
Mierzwia	2017	Poland	NEu	AHP-WLC	✓		2	✓					X	X		✓		X		✓	✓		✓					
Rediske	2020	Brazil	SAm	AHP-TOPSIS	✓	experts: 10 I, 7 A	17	✓					✓	X	✓					✓		✓	✓					5
Ruiz	2020	Indonesia	Asia	AHP	✓	D: G	5, 7	✓	✓		✓		X	X	✓	✓	✓			✓	✓		✓					5 10
Samanlioglu*	2017	Turkey	ME	fAHP- PROMETHEE			2	✓					✓											✓	✓			
Samu*	2019	Kenya	SSA				3	✓																				
Sánchez-Lozano	2013	Spain	NA/Med	AHP-TOPSIS	✓	lit. rev., D	4, 1	✓	✓				✓	X	✓	✓		✓		✓	✓	✓	✓					
Saraswat	2021	India	ISC	AHP	✓	experts	5	✓					✓	✓	✓	✓	✓			✓	✓		✓					5 40
Sindhu	2017	India	ISC	AHP-TOPSIS	✓	experts: I, A, G, E	12	✓					✓	X						✓	✓		✓	✓	✓			
Suh	2016	Korea	Asia	fAHP	✓	experts: I, A, E	10	✓	✓				X	X	✓		X		X	✓	✓							18
Surve	2014	India	ISC		✓	standard?	1	X					X	X	X				X				X			m		3
Tahri	2015	Morocco	NA/Med	AHP	✓		3	✓	✓				X	X	✓	✓				✓			✓					13
Tavana	2017	Iran	ME	fAHP-ANFIS	✓	experts	ND	✓												✓	✓		✓					
Tercan	2020	Turkey	ME	WLC	✓		5	✓					✓	✓	✓	✓	✓		✓	✓	✓		✓					13 10
Tiba	2012	Brazil	SAm				2	✓	✓			✓																
Uyan	2017	Turkey	ME	AHP	✓	experts: A	4	0					✓	X	✓					✓	✓		✓					2 5
Waewsak	2020	Thailand	Asia	AHP	✓	experts: A (I)	7	✓					✓	✓	✓		✓	X		✓	✓		✓					3 1.5
Zoghi	2017	Iran	ME	fAHP, WLC	✓	D?: I, A, G	32	✓		✓	✓		X	X	✓	✓	✓			✓	✓		✓					45
																							First Quartile		3	3		
																							Median		5	8		
																							Third Quartile		13	10		

Table 15: Abbreviations used in [Table 14](#)

	Abbreviation	Meaning
Regions	Asia	Asia excluding ISC
	ISC	Indian Subcontinent
	ME	Middle East
	NA/Med	North Africa and the Mediterranean
	NEu	Northern (i.e., non-Mediterranean) Europe
	SAm	South America
	SSA	Sub-Saharan Africa
MCDA	AHP	Analytical Hierarchy Process
	dANP	DEMATEL-based analytic network process
	fAHP	fuzzy AHP
	MCDA	Multi-Criterion Decision Making
	See specific papers for other MCDA abbreviations	

[Table 14](#) includes studies reviewed but not plotted on the weight comparison chart ([Figure 21](#)). Reasons for exclusion are discussed in [9.16](#). [Table 14](#) illustrates the large proportion of studies that involve no expert consultation, and how few studies consult industry experts. Quartile values for Slope and Population limits informed decisions about constraints to set for these variables.

Return to [2.2 Suitability studies](#); [3.1.4 Infrastructure](#)

9.3 Economic factors considered in studies reviewed

Table 16: Factors

Lead author	Year	Economic factors considered
Kim	2018	Land lease fee
Samanioglu	2017	Land cost
Sindhu	2017	Availability of skilled labour
Jun	2014	Construction costs, Operation and maintenance (O&M), GDP as proxy for demand
Kengpol	2013	Installation costs

Note that many of these factors are site-specific, and would be difficult to map for a state or region.

Return to [2.2 Suitability studies](#)

9.4 Data sources

Table 17: All data sources. All datasets obtained through web scraping or personal communication are available on request.

Category	Data/layer (note1)	Measure	Units	Date/Period	Source (reference) (note2)	Format	Resolution
Climate	Irradiance	GTI	kWh·m ⁻² ·yr ⁻¹	1999-2018	Global Solar Atlas 2.0 (SolarGis, 2019)	raster	9 arc seconds, daily (note3)
	Temperature		°C	1994-2018			
	Soiling (Dust deposition risk)	AODb	none	2013-2019	NASA MODIS (Levy <i>et al.</i> , 2015)	point (hdf)	10 km swath, daily; seasonally variable
Terrain	LULC	IGBP classification	categorical	2005	NASA ORNL DAAC (Roy <i>et al.</i> , 2016)	raster	100 m
	Elevation (→ slope)		m	2013	NASA JPL	raster	1 arc second
	Protected areas			None given	Wildlife Institute of India	raster	
Infrastructure	Roads			-	OSM via MapCruzin	line	~30 m (note4)
	Roads, K			1992	DCW via DIVA-GIS (DMA, 1992)	line	
	Powerlines			-	OSM via NextGIS	line	~30 m (note4)
Social	Population	Density	persons·gridcell ⁻¹	2020	WorldPop and CIESIN, 2018	raster	3 arc seconds
	Economic data, R	GPB	rupees	2018-2021	GoR Panchayati Raj	table (pdf)	agregated by taluk
	Economic data, K	PCI	rupees	2008-2009	GoK Directorate of Economics & Statistics (Subramanya, 2012)	table (pdf)	taluk
	Economic data, K	TDP	lakhs	2008-2009		table (pdf)	taluk
	Demand, K	Net load	MW	16/9/2020 - 30/10/2020	KPTCL	table (jpg)	Sub-station, hourly but variable
	Demand, R	Electricity billed	kWh	2018	AVVNL, JDVVNL (State electricity distribution companies)	table (pdf)	11kV feeder
SEG	Plant data, K	Location, capacity	MW	03/2019	KREDL	table (xlsx)	Village
	Plant data, R			05/2020	RSA, personal communication	table (pdf)	
Administrative boundaries	Villages, R			2001-2011	Indian Village Boundaries Project	geojson	±500 m
	Villages, K						
	All other boundaries			2015	GADM (via DIVA-GIS)	vector	

Raster, line, point and vector data accessed as GIS layers, except AODb. Abbreviations on following page. Full URLs and citations in [References](#).

Note 1: K: Karnataka; R: Rajasthan. Otherwise used for both.

Note 2: Reference included only if different from name of source

Note 3: 1 arc second ≈ 30 m at the equator

Note 4: See Lloyd *et al.*, 2017. Dates not given for OSM data due to continual updating.

Table 18: Abbreviations used in table of data sources

AVVNL	Ajmer Vidyut Vitran Nigam Ltd	JPL	Jet Propulsion Laboratory
CIESIN	Center for International Earth Science Information Network	KPTCL	Karnataka Power Transmission Corporation Ltd
		KREDL	Karnataka Renewable Energy Development Ltd
DCW	Digital Chart of the World	lakh	100,000 rupees
DMA	Defense Mapping Agency	LULC	Land Use, Land Cover
GADM	Database of Global Administrative Areas	MODIS	Moderate Resolution Imaging Spectroradiometer
GoK	Government of Karnataka	ORNL	Oak Ridge National Laboratory Distributed
GoR	Government of Rajasthan	DAAC	Active Archive Center
GPB	Grampanchayat Budget	OSM	Open Street Map
IGBP	International Geosphere-Biosphere Programme	PCI	Per Capita Income
		RSA	Rajasthan Solar Association
JDVVNL	Jodhpur Vidyut Vitran Nigam Ltd	TDP	Taluk Domestic Product

AODb, GTI and SEG as described in text.

OSM data © OpenStreetMap contributors. Licenced under the Open Data Commons Open Database Licence (ODbL), available at opendatacommons.org/licenses/odbl/.

Return to [3.1.1 SEG data](#)

9.5 Kernel Density

Kernel density can be applied to point or line data. To clarify the functioning of the kernel density tool an idealised example of application to line data is shown. [Figure 24](#) shows a map of powerline kernel density with limiting radius = 5 km. The vertical black line is a 500 kV powerline. The horizontal green line is a sampling transect. The graph in [Figure 24](#) shows powerline density values at 100 m intervals along the transect, West to East. A quartic trendline is fitted to approximate the quartic formula used to calculate kernel density. See (Silverman, 1986: 76) for further details.

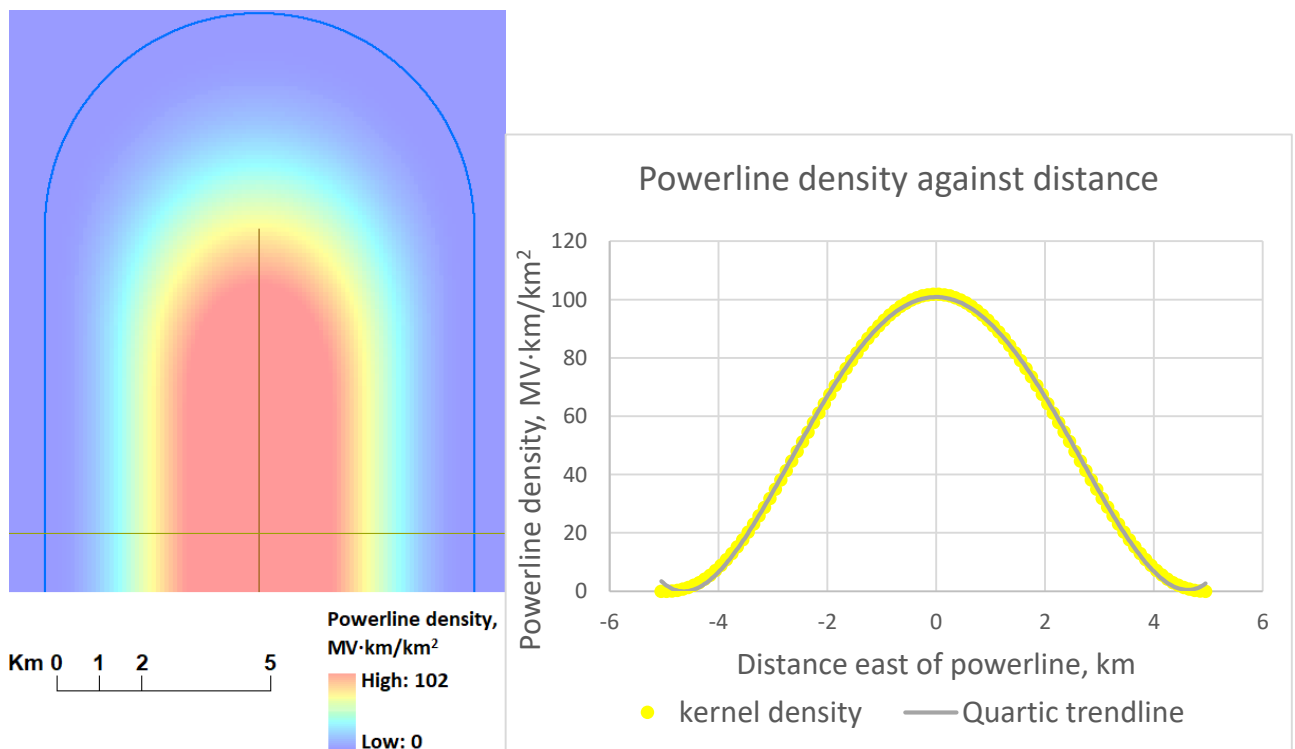


Figure 24: Illustration of kernel density function. Return to [3.1.1 SEG data](#)

9.6 AODb data code

9.6.1 Data access wget

The following code is duplicated 5-10 times and run from the command line:

```
wget -e robots=off -m -np -N -R .html,.tmp -nH --cut-dirs=3 -i "URL_list.txt" --header "Authorization: Bearer {access key}" -P "{download directory}" &
```

Where {terms} are replaced with the relevant password and local directory location respectively. URL_list.txt contains a list of target pages on the Nasa distribution server of the form https://ladsweb.modaps.eosdis.nasa.gov/archive/allData/61/MOD04_L2/2016/366/ (one page per day). The code is then pasted into the command line. The servers impose a lag on each request; duplication allows multiple requests to run in parallel, increasing overall download speed. More details on wget codes at <https://www.gnu.org/software/wget/manual/wget.html>. The basic wget code to use is available on the server archive page linked above. Wget programme authored by (Nikšić, 2015).

9.6.2 Data extraction code

The files downloaded are .hdf (Hierarchical Data Format) files. There are hundreds per day, which may or may not contain data about any given region, though the same regions tend to be surveyed at the same time each day. Each file contains data for dozens of different instruments. The following Python code inspects each .hdf file and extracts any AODb data for the study areas. Tabbing has been altered for readability.

```
"""
First, the wget scripts scrape the MODIS files from the NASA server. Then this code
extracts the AODb data from the MODIS hdf4 files and writes the data to csv
Main steps:
1) read hdf4 file
2) extract geolocation and AODb fields
3) loop thru swath: if AODb <> no data, get lat,lon
The data in csv is then processed with the separate script AODbySeason.py to gather
together data for each season
"""

import os
import csv
import rasterio as rio # this is the .hdf reader
import itertools

# function to get paths to all files in directory, including those in subfolders
def list_files(dir):
    r = []
    for root, dirs, files in os.walk(dir):
        for name in files:
            r.append(os.path.join(root, name))
    return r

# column headings. AOD_blue column is 47 nm aerosol optical depth (the variable of
interest)
headers = ["lat", "lon", "AOD_blue", "year", "day", "time"]
fileStartMarker = ".A20" # to help slice into file names and extract time stamp
scaleFactorFloat = 0.001 # scaling factor from Modis metadata, accurate to 10dp
# see article at https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4554528/
scaleFactorInt = 1 # used if int. output preferred for ease of analysis (scaleFactor
0.001 applied post-analysis)

def extractData(dir):
```



```

finalSlash = dir.rfind("\\")
csvName = dir[finalSlash + 1:] + ".csv" # name of output csv file, corresponding to
input directory
corruptFileLog = "corruptFileLog" + csvName # keep a record of no. of corrupt files
# specific corrupt files were identified and redownloaded separately
corruptFileCount = 0
with open(csvName, 'a', newline="") as csv_file:
writer = csv.writer(csv_file)
writer.writerow(headers)

dayOld = 0 # set day marker to 0
for file in list_files(dir): # iterate over all files in directory. These are the hdf4
files with AODb data.
    if not file.endswith(".hdf"):
        continue # skip over any non-data files

    fileStart = file.find(fileStartMarker) + 2 # slice in and find timestamp
    fileRef = file[fileStart:].replace(".", "") # filenames have a fairly consistent
encoding of timestamp data, but not 100 %. This standardises it.
    year = int(fileRef[:4]) # extract timestamp data
    day = int(fileRef[4:7])
    time = int(fileRef[7:11])
    # Check if we're at the start of a new day. If we are, we log info about any corrupt
files from the day we just finished scanning
    if day != dayOld: # this tells us if we're at the start of a new day
        corruptFileRecord = [dayOld, corruptFileCount] # prepare the record about corrupt files
        # log the corrupt file count to the corrupt file log
        with open(corruptFileLog, 'a', newline="") as log_file:
            logWriter = csv.writer(log_file)
            logWriter.writerow(corruptFileRecord)
        print(corruptFileRecord) # show cumulative count of corrupt files
        dayOld = day

    try:
        dataset = rio.open(file, "r") # get the whole dataset out of the hdf4 file
    except:
        corruptFileCount += 1 # if data can't be accessed, log that file as corrupted...
        continue # ...and skip to next file

    # The dataset contains all data logged by the satellite at that time and location,
including lots that is irrelevant to us. We slice in and access the specific subdatasets
that are useful.
    sdsPaths = dataset.subdatasets # get all subdatasets from within dataset
    lonsPath = sdsPaths[111] # get paths to the desired subdatasets (Longitude, latitude,
Aerosol Optical Depth (all bands)). These indices were found by inspecting the hdf4 files
with a GUI hdf viewer
    latsPath = sdsPaths[112]
    AODpath = sdsPaths[51]

    lons = rio.open(lonsPath, "r") # access longitudes
    lonsArray = lons.read(1) # read the data into an array
    lats = rio.open(latsPath, "r") # same for latitudes
    latsArray = lats.read(1)
    AOD = rio.open(AODpath, "r") # access AOD data. This contains all bands; we only want
the blue band
    AODbArray= AOD.read(1) # read band1 (47 nm AOD = blue band, AODb) into array

    # get dimensions of arrays. "Along" and "across" are the terms used to refer to the
dimensions of the swath of readings that the satellite stores in one file- a 2d path
segment. the swath has slightly different dimensions in different files- i.e., datadumps
made at different times/ places- but the same shape for all variables within a file, so
we can get the shape of any variable and use it for all of them. Choice of lons is
arbitrary.
    nAlong = lonsArray.shape[0]
    nAcross = lonsArray.shape[1]
    alongList = list(range(nAlong))

```

```

acrossList = list(range(nAcross))
# make 2d array with dimensions of combined along,across arrays- i.e, dims of swath
swath = itertools.product(alongList, acrossList)
# note: each data point has coordinates that locate it within the swath (along, across)
and *also* has coords that locate it on the globe (lat, lon). These are not the same. We
use the along, across coords to go into the arrays and get the data. The first array we
go into w the along, across coords is the AODb array. Then we use the same along, across
coords to go into the lat and lon arrays to get out the corresponding lat and lon coords.

for along, across in (swath): # iterate over pairs of along, across coords in swath
AODb_raw = AODbArray[along][across] # for each pair, get the AODb value from AODb array
if AODb_raw == -9999: # if it == the NoData value, skip to next along,across pair
    continue
# If it's not NoData, extract the data
AODb = AODb_raw * scaleFactorInt
# not relevant now, but for dev purposes it was useful to scale the AODb values
lon = lonsArray[along][across] # get lat and lon coords
lat = latsArray[along][across]
# check whether point is in Karnataka/Andhra Pradesh (AP data extracted from curiosity)
if ((74 < lon < 85) and (11.5 < lat < 19.5) or # bounding box of Karnataka + AP
(69 < lon < 78.5) and (22.5 < lat < 30.5)): # bounding box of Rajasthan
# if the data is within one of these boxes, write it to csv
row = [lat, lon, AODb, year, day, time] # prepare data
writer.writerow(row) # write data to csv for mapping

# write out the log of corrupt files
corruptFileRecord = [day, fileRef, corruptFileCount]
with open(corruptFileLog, 'a', newline='') as log_file:
logWriter = csv.writer(log_file)
logWriter.writerow(corruptFileRecord)

return (f"final corrupt file count for {csvName} at end of day {day} of year {year} =
{corruptFileCount}")

# location of input data files
# each of these directories contains numbered subfolders
# each subfolder corresponds to one day, and contains ~100 hdf4 files
# each hdf4 file contains data for one set of observations

dir1 = r"D:\UKuni\3rdYr\Project\GIS_data\AOD_trans1\modis\MOD04_L2\2013\2013ii"
...
dir9 = r"D:\UKuni\3rdYr\Project\GIS_data\AOD_trans1\modis\MOD04_L2\2014\2018ii"

dirs = [dir8, dir9] # Specifying data to access

for dir in dirs:
    try:
        extractData(dir)
    except:
        print(f"extraction failed at this point in {dir}; continuing with next dir")
        continue

```

9.6.3 AODb data collation by season

```

"""
modis_5_2.py extracts the AODb data (including lat, lon, date and time) from the hdf4
files and writes it to csv
This script takes those csvs as input and gathers data by season. This is necessary
because dust levels in the monsoon are irrelevant- any dust deposited will be washed off
almost immediately- and because some times of year were much more data-rich, so simply
averaging all datapoints would bias the output in favour of those seasons.
"""

import os
import csv

```

```

dir = r"D:\UKuni\3rdYr\Project\GIS_data\DissPy\_outputCsvs\re_in\2013" # location of
input data
outdir = r"D:\UKuni\3rdYr\Project\GIS_data\DissPy\_outputCsvs\re_in\seasons_out_K" #
where to write out
winterCsv = os.path.join(outdir, "winter.csv")
preMonsoonCsv = os.path.join(outdir, "K_preMonsoon.csv") # K_ because currently working
on Karnataka data
monsoonCsv = os.path.join(outdir, "K_monsoon.csv")
postMonsoonCsv = os.path.join(outdir, "K_postMonsoon.csv")
unassignedCsv = os.path.join(outdir, "K_unassigned.csv") # should come out empty

# open the output csvs (one for each season)
with open(winterCsv, 'a', newline="") as w,\
    open(preMonsoonCsv, 'a', newline="") as pre,\
    open(monsoonCsv, 'a', newline="") as m,\
    open(postMonsoonCsv, 'a', newline="") as post,\
    open(unassignedCsv, 'a', newline="") as u:
    # create the writer objects
    winterWriter = csv.writer(w)
    preMWriter = csv.writer(pre)
    monsoonWriter = csv.writer(m)
    postMonsoonWriter = csv.writer(post)
    unassignedWriter = csv.writer(u)
    for filename in os.listdir(dir): # iterate over AODb csvs
        fileIn = os.path.join(dir, filename)
        with open(fileIn) as csvfile: # read in data
            data = list(csv.reader(csvfile))
            for row in data[1:]:
                print("scanning row") # just so we know when something's happening
                lat = float(row[0])
                day = float(row[4])
                AODb = row[2]
                if lat > 19: # < 19 for Rajasthan
                    continue # R is all above 19, K is all below
                # select data by day of year
                if day > 334.5 or day < 59.5:
                    winterWriter.writerow(row)
                    continue
                if 59.5 < day < 151.5:
                    preMWriter.writerow(row)
                    continue
                if 151.5 < day < 273.5:
                    monsoonWriter.writerow(row)
                    continue
                if 273.5 < day < 334.5:
                    postMonsoonWriter.writerow(row)
                    continue
            unassignedWriter.append(row) # this should- and does- come out empty

```

Return to [3.1.2 Climate](#)

9.7 Karnataka electricity consumption scraper code

As for AODB data, wget scripts were used to automate file access. The Karnataka Power Transmission Company (KPTCL) website provided real-time power data at substation level for each of the five Electricity Supply Companies (ESCOMs). This appears to have been phased out as the pages are not live as of writing. The data was presented as jpegs ([Figure 25](#)). Although the KPTCL website presented these through a .aspx front end, the jpegs could be scraped directly from <http://kptclsldc.com/data1/{ESCOM}.jpg>, replacing {ESCOM} with the name of the ESCOM. The wget script below was automated to run hourly. URL_list.txt contains a list of the five {ESCOM}.jpg addresses to scrape. Jpegs were batch-cropped to leave only the relevant information and then processed with Optical Character Recognition to produce usable numerical data for analysis.

```
wget -e robots=off -m -np -R .html,.tmp -nH -i "URL_list.txt" -P {download directory}
```

57

BESCOM	MESCOM	GESCOM	HESCOM	CESC	GENERATION Admin !
Date : 13/10/2020	Time : 7:00 AM	Block : 28			
ENERGY ENTITLEMENT UPTO CURRENT BLOCK (MU) * : 6.65			NET LOAD (MW) : 612.2		
CUMMULATIVE ENERGY UPTO CURRENT BLOCK (MU)* : 1.07			NET ENT (MW) : 1051		

GESCOM LOAD MONITORING							
STATION	TRANSFORMER CAPACITY (MVA)	TRANSFORMER LOAD	OTHER INJECTIONS	NET LOAD	ENTITLEMENT	DEVIATION FROM ENTITLEMENT	ENERGY ENTITLEMENT UPTO CURRENT BLOCK (MU)
ALIPUR	2X100	108.4	15.1	123.5	93	30.5	0.59
GADAG-G	-	0.0	0	0	0	0	0
HALBURGA	2X100	64	2	66	79	-13	0.5
HAVERI-G	-	0.0	0	0	29	-29	0.19
HUMNABAD	2X100	36	0.5	36.5	66	-29.5	0.42
ITTAGI	2X100	0.8	10.4	11.2	97	-85.8	0.61
KALYANI-STEELS	EHT	35	0.0	35	35	0	0.07
KAPNOOR	2X100	78.9	5.7	84.6	97	-12.4	0.61
KUDLIGI	2X100	2.1	0.1	2.2	0	2.2	0
KUSTAGI	2X100	42.5	-0.6	41.9	75	-33.1	0.47
LINGAPUR	2X100	-7.7	17.6	9.9	81	-71.1	0.51
LINGASUGUR	2X100	-74.6	80	5.4	41	-35.6	0.26
MALLAT	2X100	19.6	0.0	19.6	0	19.6	0
RAICHUR	2X100	60.7	7.5	68.2	106	-37.8	0.67
SEDAM	2X100	0	17.3	17.3	34	-16.7	0.21
SHAHBAD	2X100	-44.6	4.2	-40.4	49	-89.4	0.32
SHAHPUR	2X100	51	2.9	53.9	73	-19.1	0.46
SINDANOOR	2X100	61	16.4	77.4	96	-18.6	0.61

Figure 25: Example power data. "Net load" is the target data.

Return to [3.1.5 Social](#)

9.8 Substation geocoding

Electricity consumption data in Karnataka and Rajasthan was obtained/aggregated at substation level. Substations are identified by place name only. The following Google Apps Script code was used to get coordinates for each place name. The list of substation names was uploaded to Google Sheets and this code called as a custom function on each.

```
function getCoordsR(address) {
  address = address + ", Rajasthan" // search google maps for "{address}, Rajasthan"
  var output = Maps.newGeocoder()
```

```

.setBounds(23, 69.5, 30.2, 78.25) // bounding box around Rajasthan
.geocode(address)
var lat = output["results"][0]["geometry"]["location"]["lat"]
var lng = output["results"][0]["geometry"]["location"]["lng"]
return writeAdjacent(lat, lng)
}
function writeAdjacent(s1,s2) { // to write in two adjacent cells with one geocode call
var results = new Array(1);
var info = new Array(2);
info[0] = s1;
info[1] = s2;
results[0] = info;
return results;
}

```

Return to [3.1.5 Social](#)

9.9 Normality of input datasets

These calculations were performed on the datasets as prepared for the regressions, such that:

- 1) Data in the exclusion zones was not counted
- 2) Population density data had been smoothed with focal statistics, radius = 10 km
- 3) Powerline density was calculated with kernel density radius = 10 km
- 4) SEG density was calculated with kernel density radius = 15 km
- 5) TDP and GPB were per-unit-area figures at taluk level

The criteria for non-normality given in (Kline, 2016: 76), reviewing a number of other studies, are that variables with $|\text{skewness}| > 3.0$ are “severely” skewed, while $|\text{kurtosis}| > 10$ suggests a problem with the data and > 20 a serious problem. Red and orange highlighting in [Table 19](#) indicate moderate and severe non-normality according to these measures. Kurtosis statistics used throughout refer to excess kurtosis.

Table 19: Skewness and Kurtosis of input datasets

Category	Variable	Karnataka		Rajasthan	
		Skewness	Kurtosis	Skewness	Kurtosis
Climate	Irradiance (GTI)	-1.20	4.61	-0.654	-0.687
	Temperature	-0.165	-0.835	-0.314	0.085
	AODb	0.233	-1.29	0.971	0.768
Terrain	Slope	3.09	19.0	1.74	4.65
Infrastructure	Roads (Euclidean distance)	2.22	6.39	1.46	3.44
	Powerline density	2.18	6.72	2.12	5.63
Social	Population density	16.9	331	10.7	206
	log(Population density)	2.08	10.1	-1.02	0.996
	PCI	6.85	55.5	N/A	
	log(PCI)	2.33	9.52		
	TDP	12.1	164		
	log(TDP)	2.45	9.95		
	GPB	N/A		11.3	154
	log(GPB)			0.383	3.88
	Electricity demand	3.41	12.1		
SEG	SEG density	25.8	794	28.0	910
	log(SEG density)	-1.33	3.39	-0.446	1.36

Demand was not transformed due to the narrow margin by which it exceeded the recommended limits. The untransformed datasets of all other social variables, and SEG density, display severe skewness and kurtosis, while their log transforms do not. The possible exception to this is the kurtosis of log(Population density) in Karnataka, which remains marginally problematic. Given the lack of consensus on cut-offs for these statistical measures (see discussion in e.g. Byrne, 2016), such marginal cases were considered acceptable.

Return to [3.2 Data transformations](#)

9.10 Distance to roads map layer, Rajasthan

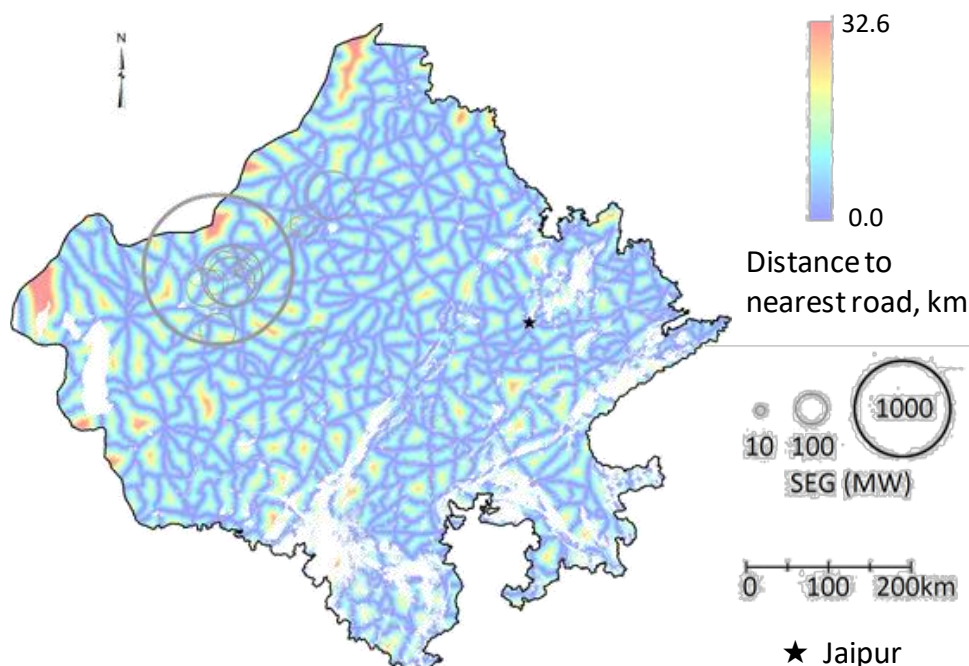


Figure 26: Euclidean distance to roads. Return to [Figure 7](#)

9.11 Matlab Multilinear Regression code

```
% Purpose of code: Run backwards stepwise regressions on map data.
% Dependent variable is SEG density (or log (SEG density)). Independent
% variables are other map layers (temperature, solar irradiance (GTI),
% log(per capita income), etc.)

% read in data as table
nonlogfile =
"D:\UKuni\3rdYr\Project\GIS_data\DissPy\Regression\ascii_reg_output\K\ML_input\A
RO_l10s_v5_TL.csv";
logfile =
"D:\UKuni\3rdYr\Project\GIS_data\DissPy\Regression\ascii_reg_output\K\ML_input\A
RO_logPVo_l10sFS40_v4.csv";
ktkStats = readtable(nonlogfile);

% Set segment to N for North, S for South or A for All.
% (Karnataka only; N/S segmentation not used for Rajasthan)
% "All" is assumed by default, but it's still useful to have "A" written on
% the output data.
% The data are originally exported from the map as a grid, then the rows
% concatenated and transposed, so the top half of the table approximates to
% the northern half of the map. For the SEG density data the midpoint is
% 6.7 km north of the exact north-south centre line. For the log density
% data it's a further 11.5 km north. This 11.5 km difference is <1.5 % the
% NS length of the state.
% Each row contains data for one cell on the map; each column is a
% different variable
```

```

segment = "s"

% specify size of subsample to be taken for partial analyses
tdsample = 0.35
% Regressions were run on subsets to see how consistent the coefficients
% were. This was helpful- coefficients that were more variable across
% subsets did turn out to be less predictive- but the approach was
% superseded by the MWR code, so results of partial analysis are not
% discussed in the report.

% get number of rows in table
nrows = height(ktkStats);
% get index of middle row
midrow_idx = round(nrows/2);

% if only the north or south half of the table is desired:
if ismember(segment,["N" "n"])
    ktkStats = ktkStats(1:midrow_idx,:); % if north, take table from top
                                         % to middle
    nrows = height(ktkStats);
elseif ismember(segment,["S" "s"])
    ktkStats = ktkStats(midrow_idx+1:end,:); % if south, middle to end
    nrows = height(ktkStats);
end

% Regression
% first, run backwards stepwise linear regression on full data set
% column names are: pv gti aodb t slope load pl_dens l10pci l10tdp ed_rds
% logpopfs10...20...30...40 logpopfsan1020...2030...3040

% define model
mdl = stepwiselm(ktkStats, 'pv~gti+aodb+slope+pl_dens+logpopfs20',...
    'Upper','linear', 'ResponseVar','pv','PredictorVars',...
    {'gti','aodb','slope','pl_dens','logpopfs20'})

% second, for partial analysis, run on 100 random subsets
% get number of rows that corresponds to a chosen %age of the full sample
% (or of specified half if doing segment analysis)
rows_to_sample = round(tdsample * nrows);
nloop = 0; % initialise loop counter
for nloops = 1:100 % for 100 loops...
    nloop = nloop + 1 % increment visible loop counter
    % pick a random subsample of rows from the table (size of sample is
    % specified above in tdsample)
    idxs = randperm(nrows, rows_to_sample);

    % take those rows from the table
    ktkStatSample = ktkStats(idxs,:);

    % run a backwards stepwise linear model on those rows
    mdl = stepwiselm(ktkStats, 'pv~gti+aodb+slope+pl_dens+logpopfs20',...
        'Upper','linear', 'ResponseVar','pv','PredictorVars',...
        {'gti','aodb','slope','pl_dens','logpopfs20'})
end
% Regression data (coefficients etc.) are then exported to Excel for
% collation, analysis and comparison
See Figure 27 for e.g. output.

```

```

Editor - D:\UKun\3rdYr\Project\GIS_data\DisPy\Regression\ascii_reg_clean1.m
51 % column names are: pv gti aodb t slope load pl_dens l10pci l10tdp ed_rds
52 % logpopfs10...20...30...40 logpopfsan1020...2030...3040
53
54 % define model
55 mdl = stepwiselm(ktkStats, 'pv~gti+aodb+slope+pl_dens+logpopfs20',...
56 'Upper','linear','ResponseVar','pv','PredictorVars',...
57 {'gti','aodb','slope','pl_dens','logpopfs20'})
58
59 % second, for partial analysis, run on 100 random subsets
60 % get number of rows that corresponds to a chosen size of the full sample

Workspace
+
ascii_reg_clean1.m
GDR_2_0.m
r_ascii_reg_1.m
getlowvals.m
getlowvalsTG2.m

Command Window
>> ascii_reg_clean1

segment =

    "s"

tdsample =

    0.3500

No terms to add to or remove from initial model.

mdl =

Linear regression model:
    pv ~ 1 + gti + aodb + slope + pl_dens + logpopfs20

Estimated Coefficients:

               Estimate          SE          tStat          pValue
               _____          _____          _____          _____
(Intercept)    -1.8514         0.033414    -55.409           0
gti             0.00094672     1.5817e-05     59.854           0
aodb           -0.052564         0.0084705    -6.2056     5.4565e-10
slope          -0.0022506       0.00022235   -10.122     4.4476e-24
pl_dens         4.7679e-07      1.3783e-08    34.593     1.2183e-261
logpopfs20     -0.15665         0.0025038   -62.566           0

Number of observations: 269126, Error degrees of freedom: 269120
Root Mean Squared Error: 0.256
R-squared: 0.0334, Adjusted R-Squared 0.0334
F-statistic vs. constant model: 1.86e+03, p-value = 0
fx >>

```

Figure 27: output of above code (selected non-log regression for South Karnataka)

Return to [3.4 Regression testing](#)

9.12 Matlab Moving Window Regression code

9.12.1 Driver code (front end)

```

%%%%%%%%
% Purpose: assess spatial variation in regression coefficients.
% Mechanism:
% MWR (Moving Window Regression) is like a simpler version of GWR
% (Geographically Weighted Regression). It takes map data for a dependent
% variable Y and some number of independent variables X1, X2...Xn. Data
% format is described below. The arrays in have the shape of the map data,
% i.e. each array could be mapped directly onto the map using its i,j
% coordinates in the array as lon and lat (with appropriate scaling and
% translation constants).
% User specifies a bandwidth bw and a spacing sp. The code then scrolls
% through the Y and X1...Xn arrays, starting in the top left corner and taking
% for each variable a square subarray corresponding to a map "window" of
% edge length bw. It runs a linear regression on these subarrays and writes
% the regression data (r2, pvalue, coefficients etc.) out to csv. It then
% moves to the right by a distance equal to the specified spacing and takes
% a new set of subarrays (the next window). It continues along the row
% until it reaches the end, then moves down by sp and starts again at the
% beginning of the next row.

```

```

#####

%%% USER INPUT BLOCK %%%
% enter file path info for input/output data
path =
'D:\UKuni\3rdYr\Project\GIS_data\DissPy\Regression\resampled_asciis\snap_15_500_
full\';

% filenamearray is array of input file names
% Inputs are .txt files written out from ArcMap with Raster to ASCII tool
% First one needs to be the y vals (i.e., SEG or log(SEG) vals)
% Rest are x vals (independent variables)
% Data is set up such that the filenames have the form variablename +
% batchtag + .txt, so the filename without batchtag can be used for
% labelling columns in the output data
% The filename array contains all the possible variables. For each run,
% variables not to be included are commented out. So this run modelled
% log(SEG density) in terms of GTI, AODb, slope, powerline density and
% (log(population density))_(averaged at 30km)
filenamearray = {
%     'pv' % original var name "PV", not "SEG". Check back-compatibility
    'L10pv'
    'GTI'
    'aodb'
%     'T'
    'slope'
%     'load'
    'pl_dens'
    'L10PCI'
%     'L10TDP'
%     'ed_rds_dv'
%     'logpop_fs20'
    'logpop_fs30'
};

batchtag = '_15_500_full.txt'; % invariant for each dataset
ncheckval = 2; % which of the vars you want exported as a checkval
                % to doublecheck the imported points are in the right place
                % on the map. Y is 1, X1 is 2, Xn is n+1

% path to write the data out
outpath =
'D:\UKuni\3rdYr\Project\GIS_data\DissPy\Regression\resampled_asciis\snap_15_500_
full\GDR_logpv';
% first bit of the outfile name, specified by user at each run to identify
% output files:
outfilestem = 'logPVb_GASPldlPcilPFS30';
% outfiletag could just be .csv, or can include an additional version
% identifier ('b.csv'; 'version_2.csv') if running slightly modified
% versions of the same basic regression
outfiletag = '_clchk1.csv';
% dimensions, in meters, of map cells:
[cellx, celly] = deal(500,500);
% coords of top left corner of map (get from Arc):
[refx, refy] = deal(398334.8956, 2032383.292);

% we're going to create subarrays by slicing into our main arrays
% first we specify the bandwidth (edge length) of our subarrays, in km:
bw = 300;
% next we specify the spacing, in km, that we want between our focal points
% (the subarray centres). If sp = bw, there will be no overlap, and each
% cell will be counted only once. sp << bw -> lots of overlap -> focus will
% move more slowly over the map, producing a more smoothly varying output.
sp = 40;

% Now we set the maximum fraction (0-1) of a subarray that can be NaN

```

```

% before we reject the subarray. Tried playing with this but for non-log
% SEG 0.5 seems best. Otherwise you get centres plotted that are off the
% edge of the mapped data.
% With logPV data it needs to be higher though- there's gaps in the data bc
% all 0 vals -> NoData. NaNmax = 0.7 works at bw = 150. Presumably the
% further one zooms in the lower NaNmax can be. 0.6 works for bw = 150.
NaNmax = 0.7;

##### END OF USER INPUT BLOCK #####

t_start = now; % timer. Used in dev to compare speed of diff approaches.

##### READ BLOCK #####
dimsfilenamearray = size(filenamearray); % dimensions of array of filenames
nvars = dimsfilenamearray(1); % n variables (incl. y) = n files in array
YXarr = {}; % empty cell array to populate w data
for n = 1 : nvars
    filename = filenamearray{n}; % get nth filename
    filename = strcat(filename, batchtag); % append file suffix
    vals = dlmread(fullfile(path, filename)); % read data
    YXarr{n} = vals; % add data to array
end

% YXarr now holds all the data. YXarr has n cells at the top level, where n
% = nvars = number of variables (Y and X). Each one of those cells
% contains an i by j array, where i and j are the dimensions of the map (in
% gridcells, not metres)

##### RUN BLOCK #####
% First we make the output filepath using the kilometre measures of bw, sp
% This makes it easier to see which data is from which run
outname = [outfilestem, '_bw', num2str(bw), 's', num2str(sp), outfiletag];
outfile = fullfile(outpath, outname);

% Create a csv with this name and write out the column headers
% makeheaders is a helper function that concatenates the variable name with
% the statistic type (coefficient, p value etc.) to make the headers for
% the output csv. It also writes the fixed headers for model stats.
% Not shown because administrative.
makeheaders(filenamearray, ncheckval, outfile);

% next we convert bandwidth and spacing from km to number of cells:
bw = bw * 1000 / cellx; % cellx = celly, so can use either
sp = sp * 1000 / cellx;

% Next we see how many subarrays will fit into our data arrays on each axis
% This tells how many rows and columns of subarrays we'll have
% Any data points in a strip of width bw-1 cells along the right and bottom
% of the main arrays will be ignored. In practice, the impact of this will
% be very small- the dataset on the map is an irregular shape, so most of
% the edge of the map is NoData anyway.

[gridY, gridX] = size(YXarr{1}); % dimensions of the data arrays
n_cols = floor((gridX-bw)/sp)+1; % n subarrays in x direction
n_rows = floor((gridY-bw)/sp)+1; % n subarrays in y direction
% Note that for these n subarray calcs it's necessary first to subtract the
% bandwidth from the grid dimensions- otherwise, when the top left of the
% subarray gets close to the bottom/right of the map, the bottom/right of
% the focal window is off the edge.

% Now we make counters to help us iterate
i = 1; % counter for x direction
j = 1; % counter for y direction
while j <= n_rows % go row by row till we hit the target number of rows

```



```

suby_top = (j-1)*sp + 1;           % get index of top and bottom of row
suby_bottom = suby_top + bw - 1;
while i <= n_cols % go col by col up to target number of cols in row
    subx_left = (i-1)* sp + 1; % get left and right limits of column
    subx_right = subx_left + bw - 1;
    % get subarrays from each main array:
    subarrs = {nvars}; % make cell array to hold subarrays
    for n = 1:nvars
        vararr = YXarr{n}; % get main array for 1st,2nd,...,nth variable
        % index into it to get subarray and add that to the cell array:
        subarrs{n} = vararr(suby_top:suby_bottom,subx_left:subx_right);
    end

% Before processing the subarrays to run the regression, we check
% how many NoData values (-9999) there are. Because the map data is
% preprocessed to make all arrays perfectly coterminous, we only
% have to check one, except when using log SEG data- that array is
% more limited, as explained above, so for log runs we have to use
% the first (Y) array.
nNoData = sum(subarrs{1}(:) == -9999); % number of NoData vals
if nNoData > bw * bw * NaNmax          % acceptable limit
    i = i + 1;
    continue
end
% If the window was too NoData-heavy, it will now have been
% discounted. If it's still being processed, we know there aren't
% too many NoData values, but there may still be some, so we check.
% If we find any, we replace them with NaN in all subarrs:
if nNoData > 0                        % check for NoData vals
    for n = 1:nvars
        subarr = subarrs{n};          % get the nth subarray
        subarr(subarr==-9999)=NaN;     % set -9999 vals to NaN
        subarrs{n} = subarr;          % put the subarray back in
    end
end
% Now we turn our subarray into a vertical array with one column
% per variable and one row per mapcell
% stackcols takes an n by m array and stacks all the columns on top
% of each other, giving an nm by 1 array. We do this with the
% subarray of Y data first:
subarrs_in = stackcols(subarrs{1});   % array with col of Y vals
% and then for each remaining var...
for n = 2:nvars
    subarr = subarrs{n};               % ...we take the subarray...
    vsubarr = stackcols(subarr);        % ...stack it vertically...
    subarrs_in = horzcat(subarrs_in, vsubarr); % ...add col to array
end
% subarrs_in is now nm high and nvars wide

% Now we get the lat and lon coords of the centre of our subarray
% so we can locate it on the map again. Mean of edge vals gets us
% the centre; timesing by cell dims converts to distance on map.
lon = refx + (mean([subx_left, subx_right]) - 1) * cellx;
lat = refy - (mean([suby_top, suby_bottom]) - 1) * celly;

% checkval is a reference value used during dev to check the mapped
% values are lining up correctly. If they are, this reference value
% should be the same as the original value in the layer on the
% map in the cell where the generated datapoint lands.
ref_in_sub = floor(bw/2);
subtocheck = subarrs{ncheckval};
checkval = subtocheck(ref_in_sub,ref_in_sub);

% Now we run the regression on our selected data by calling the
% regressout function that runs the regression and writes the data
% to csv for mapping (see below)

```

```

        regressout(checkval, lon, lat, subarrs_in, outfile);
        i = i + 1; % increment x counter to move to next column within row
    end
    % When above loop finishes, we're done with one row & move to the next
    j = j + 1; % increment y counter to move to next row
    i = 1;      % reset x counter to move to beginning of row
end
t_end = now;
% print how long it took (86400 converts from days to seconds)
runtime = (t_end-t_start) * 86400

```

9.12.2 Regression code (back end)

% This is the backend MWR code: the bit that actually runs the regression
 % on the data selected by the driver code, and writes the output to csv.
 % This code is run on each window.

```

function [data_out] = regressout(checkval, i, j, subarrs_in, outfile)
% subarrs_in contains all the input data to be regressed

pthresh = 0.05; % critical p value
Ys = subarrs_in(:,1); % first column is Y vals
dims_subarrs_in = size(subarrs_in); % get dimensions
ncols = dims_subarrs_in(2); % get number of columns
Xs = subarrs_in(:,2:ncols); % remaining columns are X1, X2... Xn vals

% run regression on Xs and Y
[b,se,pval,inmodel,stats] = stepwisefit(Xs, Ys);
% Each of these outputs is an array:
% coefficients, standard errors, p value, input (which we ignore), model
% stats

% Make an array to hold the data we're going to write out:
nxs = ncols - 1; % 1 y col, so the rest are x cols
ncols_out = 11 + 8*nxs; % 11 cols of model data + 8 cols for each x var
% Make array of NaNs with this many columns and one row (each window will
% be described by one row in the output csv)
data_out = NaN(1,ncols_out);

% Write out coords and checkval
% Coords are in metres, calculated to coincide with Arcmap's values
% Grid system used here is Kalianpur 1975 (UTM 43N). See driver code for
% explanation of checkval
data_out(1) = i; % lon
data_out(2) = j; % lat
data_out(3) = checkval;

% Now we want to write out the model stats for this subarray regression,
% starting with the general model stats. We'll do stats for each ind var
% later.
% root mean square error:
rmse = stats.rmse;
data_out(4) = rmse;

% r^2 and adjusted r^2
% source of formulae is https://uk.mathworks.com/matlabcentral/answers/93200-how-can-i-obtain-the-r-squared-and-adjusted-r-squared-values-from-stepwisefit-in-the-statistics-tool
nY = length(Ys);
varY = var(Ys, 'omitnan');
degfree = stats.df0;
r2 = 1 - (rmse^2/varY) * ((nY-1-degfree)/(nY-1));
adjr2 = 1 - rmse^2/varY;
data_out(5) = r2;

```

```

data_out(6) = adjr2;

% Assorted other model stats
data_out(7) = stats.fstat;      % F statistic
data_out(8) = stats.pval;      % p value
data_out(9) = stats.SSresid;    % sum of squares of residuals

% Calculate and write out mean and stdev of Y data
meanY = mean(Ys, 'omitnan');
stdevY = std(Ys, 'omitnan');
data_out(10) = meanY;
data_out(11) = stdevY;

% Now we loop through the model info for the independent variables
% For the nth X variable, we want the nth member of each array of model
% stats
for n = 1:nxs
    xvals = Xs(:,n);
    % first 11 cols are holding general model info, so we start writing at
    % column 12
    N = 12 + 8*(n-1);

    data_out(N) = mean(xvals, 'omitnan'); % mean of Xn in window
    stdevX = std(xvals, 'omitnan');
    data_out(N+1) = stdevX;               % standard deviation of Xn in window

    pvalue = pval(n);
    data_out(N+5) = pvalue;
    % if the pvalue is greater than threshold, the ind var will not be
    % included in the final model, and we don't want to plot info about
    % it
    if pvalue > pthresh
        continue
    end
    data_out(N+2) = b(n);                 % estimated coefficient
    data_out(N+3) = se(n);                 % standard error
    data_out(N+4) = stats.TSTAT(n); % T statistic
    bstar = b(n) * stdevX/stdevY; % standardised regression coefficient
    data_out(N+6) = bstar;
    data_out(N+7) = abs(bstar);           % abs of b* for comparing magnitudes
end
% write data out to csv:
dlmwrite(outfile,data_out,'delimiter',' ','-append','precision',7);

```

9.12.3 Helper function

```

function [column] = stackcols(arrayIn)
% takes array and stacks all the columns on top of each other to make one
% vertical column
% makes it easier to run regressions
column = reshape(arrayIn,[],1);
end

```

Return to [3.4 Regression testing](#)

9.13 Regression Equations

Table 20: Regression equations and corresponding β^* values

Run		Regression Equation	Standardised regression coefficients β^* , %							
			Climate		Terrain	IS		Social		
			GTI	AODb	Slope	PLD	Rds	Pop ₁₅	Pop ₁₅₂₅	Pop ₂₅₃₅ GPB
Rajasthan	Non-log	Control	-0.367+(1.69E-4)·GTI-(6.40E-4)·Slope+(5.36E-7)·PLD	1.0		-0.1	1.9			
		R-Pop	-0.104+(5.50E-5)·GTI-0.0412·AODb-0.00172·Slope+(6.51E-7)·PLD-0.114·Pop ₁₅ +0.184·Pop ₁₅₂₅ -0.114·Pop ₂₅₃₅	2.0	-1.5	-1.1	11.1		-22.9	35.4 -21.6
		R-GPB	-0.121+(4.28E-5)·GTI-0.00131·Slope+(5.78E-7)·PLD-0.0602·GPB	1.6		-0.8	9.9			-8.8
	Log	R-Pop	-3.53+(7.74E-4)·GTI-1.54·AODb-0.0314·Slope+(3.81E-6)·PLD-(1.54E-5)·Rds-0.428·Pop ₁₅ +0.764·Pop ₁₅₂₅ -1.29·Pop ₂₅₃₅	4.6	-8.8	-3.4	13.7	-4.1	-13.9	23.9 -40.1
		R-GPB	-3.53+(2.83E-4)·GTI-0.878·AODb-0.0160·Slope+(3.52E-6)·PLD-(4.18E-6)·Rds-1.38·GPB	1.7	-5.0	-1.7	12.6	-1.1		-32.9
Karnataka	Non-log		GTI	AODb	Slope	PLD		Pop ₂₀	Pop ₃₀	Demand PCI
		Control	-1.22+(6.06E-4)·GTI-0.00142·Slope+(1.41E-7)·PLD	9.5		-1.6	2.8			
		Global	-1.22+(6.42E-4)·GTI-0.0942·AODb-0.00222·Slope+(2.27E-7)·PLD-0.10635·Pop ₂₀	10.0	-5.6	-2.4	4.5		-9.9	
		North†	0.302-0.0631·AODb-(7.64E-4)·Slope+(8.29E-8)·PLD+(2.15E-4)·Demand-0.0555·PCI		-8.7	-2.2	5.2			2.7 -12.0
		North‡	0.0460-0.0428·AODb-(8.72E-4)·Slope+(4.06E-8)·PLD-0.00339·Pop ₂₀ +(2.40E-4)·Demand		-5.9	-2.6	2.6		-0.7	3.0
	Log	South	-1.85+(9.47E-4)·GTI-0.0526·AODb-0.00225·Slope+(4.77E-7)·PLD-0.157·Pop ₂₀	12.6	-1.3	-2.0	6.9		-12.8	
		Global	-7.67+0.00315·GTI-0.414·AODb-0.0179·Slope-1.55·Pop ₃₀	7.6	-4.9	-3.4			-21.1	
		North†	2.31-1.55·AODb-0.0134·Slope+(1.54E-6)·PLD-0.774·PCI		-12.8	-2.4	5.9			-8.4
		North‡	-1.13-1.51·AODb-0.0140·Slope+(1.06E-6)·PLD-0.111·Pop ₃₀		-12.4	-2.5	4.1		-1.1	
		South	-10.7+0.00478·GTI-1.70·AODb-0.0161·Slope-1.84·Pop ₃₀	12.3	-11.3	-3.2			-28.9	

†Northern regressions using PCI (log per capita income) were used for [Mapping and evaluation](#).

‡Northern regressions using Population variables were plotted in [Figure 21](#) for comparison with other studies, to better present the range of Population weights. Note that other β^* values are broadly consistent between northern runs despite the change of social variable. Note also negative β^*_{Pop20} in northern non-log regression, contradicting the positive values found in the MWR ([Figure 13](#)).

GTI = Irradiance, PLD = Powerline Density, Rds = Roads (Euclidean distance). Other abbreviations as in report, i.e. GPB = log(areal GPB), etc. Note non-comparability of social variables between states.

Equations presented in [Table 20](#) are those produced in Matlab for the selected models (see [Figure 27](#)), and which were used to produce the suitability maps by replacing the variable names with corresponding layer names in ArcMap's raster calculator.

Return to: [4.1.2 Forwards phase](#) (Rajasthan); [Figure 21](#) Weights comparison plot.

9.14 Normality and homoscedasticity of residuals

Table 21: Skewness, kurtosis and Breusch-Pagan heteroscedasticity of model residuals

State	Regression type	Run	Normality		Heteroscedasticity	
			Skewness	Kurtosis	F statistic	P value
Rajasthan	Non-log	Control	27.7	901	2268	0
		R-Pop	27.6	896	1591	0
		R-GPB	27.7	900	2236	0
	Log	R-Pop	-0.806	1.72	<div></div> 310	0
		R-GPB	-0.848	1.82	<div></div> 484	0
Karnataka	Non-log	Control	25.9	797	<div></div> 381	3.3E-247
		Global	25.8	794	<div></div> 547	0
		North	3.36	15.8	<div></div> 348	0
		South	19.6	437	<div></div> 369	0
	Log	Global	-1.43	3.79	<div></div> 130	3E-111
		North	-1.57	4.46	<div></div> 98	1.41E-83
		South	-1.40	3.65	<div></div> 93	6.48E-79
Dharshing, 2017					<div></div> 342	0
Balta-Ozkan <i>et al.</i> , 2015					<div></div> 35	0

Note pronounced non-normality indicated in [Table 21](#) for all non-log regressions, but none of the log regressions. The Breusch-Pagan test was selected for its computational efficiency, due to the large datasets. Within each state, the Breusch-Pagan F statistic, a quantitative measure of heteroscedasticity, is lower for log than non-log regressions. For both log and non-log regressions the F statistics are lower in Karnataka than Rajasthan. Karnataka's regional log regressions have the highest p values and lowest F statistics, indicating lower heteroscedasticity. For comparison, figures are included for the only two papers reviewed for this study that quantified heteroscedasticity of model residuals. P values not directly comparable due to rounding in source papers.

Return to [4.1.3 Mapping and Evaluation](#) (Rajasthan); [4.2.4 Mapping and evaluation](#) (Karnataka); [5.2 Method Evaluation](#)

9.15 MWR correlation strengths

Correlation strengths were calculated for the outputs of the Moving Window Regressions ([Table 22](#)). Five types of relationship were analysed for each variable in each regression:

- 1) Correlation of the variable's standardised regression coefficient β^* with latitude
- 2) Correlation of the variable's regression coefficient β with latitude
- 3,4) Correlation of β^* and β with the value of the variable itself
- 5) Correlation of the variable with latitude

While (1) and (3) are of most interest in terms of how each variable's apparent importance as an SEG siting criterion is explained, (2) is necessary to evaluate the legitimacy of the assumption that coefficient value varies somewhat smoothly with latitude, as this underlies the equation used for HSTR.

Table 22: Correlation strengths

Category	Variable	Correlation strength between outputs of MWR (Pearson's R ²)									
		Non-log regressions					Log regressions				
		Coefficients				Value vs latitude	Coefficients				Value vs latitude
		Vs latitude		Vs value			Vs latitude		Vs value		
		β*	β	β*	β		β*	β	β*	β	
Climate	Irradiance	0.27	0.22	0.04	0.04	0.02	0.35	0.34	0.11	0.14	0.03
	AODb	0.08	0.07	0.02	0.04	0.89	0.23	0.15	0.23	0.13	0.91
	Temperature	0.41	0.27	0.35	0.21	0.86					
Terrain	Slope	0.03	0.09	0.04	0.03	0.82	0.25	0.19	0.20	0.14	0.83
IS	Powerlines	0.07	0.05	0.06	0.11	0.53	0.62	0.56	0.72	0.72	0.74
Social	Demand	0.24	0.19	0.02	0.02	0.57					
	Population	0.58	0.19	0.28	0.02	0.33	0.74	0.47	0.16	0.03	0.33
	PCI	0.01	0.07	0.03	0.07	0.67	0.40	0.43	0.33	0.34	0.68
Mean		0.21	0.14	0.10	0.07	0.59	0.43	0.36	0.29	0.25	0.59
Median		0.16	0.14	0.04	0.04	0.62	0.38	0.38	0.21	0.14	0.71

Return to [4.2.3.1 Non-log MWR](#)

9.16 Weight standardisation for comparison with previous studies

Studies reviewed consider different variables and use different weighting scales. This process converted unstandardised weights w to standardised weights w^* for comparison and plotting. A set of 7 standard variables was chosen: all of the criteria evaluated in the report, with the exception of economic variables TDP, PCI and GPB as these are not considered in this form by any other study reviewed. Studies were not included here if they do not give numerical weights, if they evaluate < 3 of the standard variables, or, in the case of hybrid wind/solar studies, if separate weights are not given for each.

If a study does not discuss a variable, that variable is not assigned 0 weight, as the study has not taken any position on it. By contrast, some studies explicitly assign 0 weight to variables, and others discuss variables and reject them as unimportant, which was also considered as 0 weight for the following calculations.

As studies consider different numbers of standard variables, it is not enough to set the total standard weight $\sum w^*$ to 100 % for all studies: if this is done, $|\bar{w}^*|$ (mean absolute w^*) in a study that considers all 7 standard variables = (100 %)/7 \approx 14%, while in a study that considers only 3 standard variables $|\bar{w}^*| \approx$ 33 %. It is necessary to use absolute values to calculate total unstandardised weight because population weights are signed.

To ensure consistent \bar{w}^* , $\sum w^*$ for each study s was set as shown in [Equation 10](#)

$$\sum w_s^* = \frac{n_s}{n^*} \cdot 100\%$$

Equation 10

Where n_s is the number of standard variables evaluated by the study and n^* is the total number of standard variables (= 7).

In each study s the standardised weight of each standard variable $x_{s,1} \dots x_{s,n_s}$ is therefore:

$$w_{s,n}^* = w_{s,n} \cdot \frac{\frac{n_s}{n^*} \cdot 100\%}{\sum_{n=1}^{n_s} |w_{s,n}|}$$

Equation 11

I.e., the unstandardised weight multiplied by the ratio of the intended total standardised weight to the total unstandardised weight.

To calculate standardised weights for the suitability indices generated from the regressions used in this study, this standardisation procedure was applied to the standardised regression coefficients β^* .

Return to [Figure 21](#) Weights comparison plot.