Weatherman: Exposing Weather-based Privacy Threats in Big Energy Data

Dong Chen, and David Irwin
University of Massachusetts Amherst
{dongchen,deirwin}@umass.edu

***Abstract*—Smart energy meters record electricity consumption and generation at fine-grained intervals, and are among the most widely deployed sensors in the world. Energy data embeds detailed information about a building's energy-efficiency, as well as the behavior of its occupants, which academia and industry are actively working to extract. In many cases, either inadvertently or by design, these third-parties only have access to *anonymous* energy data without an associated location. The location of energy data is highly useful and highly sensitive information: it can provide important contextual information to improve big data analytics or interpret their results, but it can also enable third-parties to link private behavior derived from energy data with a particular location. In this paper, we present Weatherman, which leverages a suite of analytics techniques to localize the source of anonymous energy data.**

Our key insight is that energy consumption data, as well as wind and solar generation data, largely correlates with weather, e.g., temperature, wind speed, and cloud cover, and that every location on Earth has a distinct *weather signature* that uniquely identifies it. Weatherman represents a serious privacy threat, but also a potentially useful tool for researchers working with anonymous smart meter data. We evaluate Weatherman's potential in both areas by localizing data from over one hundred smart meters using a weather database that includes data from over 35,000 locations. Our results show that Weatherman localizes coarse (one-hour resolution) energy consumption, wind, and solar data to within 16.68km, 9.84km, and 5.12km, respectively, on average, which is *more accurate* using *much coarser resolution data* than prior work on localizing *only anonymous solar data* using solar signatures.

I. Introduction

Smart energy meters, which measure and transmit electricity usage at fine-grained intervals, e.g., every hour or less, are being widely deployed by utilities in many parts of the world. Smart meter penetration is expected to increase as utilities continue to upgrade old meters and support more sophisticated smart grid functionality. In addition to smart meters, end-users are also increasingly deploying Internet-enabled meters to track their own local energy consumption and generation. Renewable solar and wind installations typically include such end-user metering by default.

Given the scale of the deployments above, developing techniques that analyze big energy data to improve energy-efficiency has become an active research area in both industry and academia. Numerous startups, including Bidgely [1], Onzo [2], and Sense [3], are now focused on monetizing insights drawn from big energy data. These insights have the potential to significantly improve energy-efficiency at massive scales, e.g., by providing real-time energy-efficiency recommendations to users or automatically identifying faults in individual buildings or the electric grid. To gain these insights, utilities often contract with the third-party energy data analytics companies above and directly provide them energy meter data, while end-users often link their meters to public APIs that allow analytics companies to directly access their energy data. These third-party companies often provide analytics services to end-users "for free," since the energy data provides value to them, e.g., either as training data or in profiling users' energy usage and behavior.

Importantly, the energy data made available to the third-party companies and academic researchers above is often *anonymous* and not associated with a specific location. Anonymous energy data includes only a series of tuples, which each specify a timestamp and energy consumption (or generation). The primary reason for anonymizing energy data is to prevent third-parties from linking private information derived from energy data with a particular location [4]. However, such private information is potentially valuable. As one example, analyzing energy data can reveal irregular sleeping patterns, e.g., based on sporadic energy usage at night, which pharmaceutical companies could use to inform direct marketing campaigns of insomnia drugs. To guard against these privacy threats, the Voluntary Code of Conduct (VCC) for managing customer energy data recently released by the DOE recommends utilities remove any identifying information from energy data they share with third-parties [5].

In this paper, we present Weatherman, a suite of big data analytics techniques that extract location from anonymous energy consumption, wind, and solar data. Our key insight is that energy data largely correlates with the local weather, e.g., temperature, wind speed, and cloud cover, and that every location on Earth has a distinct *weather signature* that uniquely identifies it. Weatherman leverages this insight to localize the source of anonymous energy data. To do so, Weatherman combines physical system models with statistical techniques to extract a weather signature from energy data at each location when searching a massive weather database that includes records from 35,000 locations.

Our goal is to explore the severity of this privacy threat by quantifying the localization accuracy for energy consumption, wind, and solar data. Based on the DOE's VCC, users often do not consider the privacy implications of releasing anonymous energy data to third-parties, assuming the data is anonymous if it is not associated with location information, e.g., an address. Understanding the localization threat is important in i) educating users about the sensitivity of energy data, ii) informing evolving policies on managing energy data, and iii) developing techniques that preserve privacy, while also enabling well-intentioned analytics. Existing techniques for preserving privacy in energy data do not consider localization threats, and thus cannot prevent them [6], [7]. Broadly, Weatherman shows how public access to large "big data" archives of sensor data can introduce serious privacy threats. Our hypothesis is that weather-based localization of energy consumption, wind, and solar data is accurate to a small region. Since wind and solar sites are identifiable via public satellite imagery within the region [8], [9], such localization represents a serious privacy threat, as it may be possible to associate data with a specific home. In evaluating our hypothesis, we make the following contributions.

Weather-based Energy Modeling. We present physical models that characterize the energy consumption of buildings and the energy generation of wind and solar sites based on the weather. These physical models show how energy consumption, wind, and solar energy data correlate with specific weather metrics—temperature, wind speed, and cloud cover—in different ways, which enables localization by correlating the energy data with weather data.

Weather-based Energy Localization. We combine our physical models with statistical techniques to extract a unique weather signature at each possible location from each type of energy meter data. Weather-based localization then involves searching a massive weather database to find a location with weather that best matches the weather signature. Given the scale of the database, a key challenge is making this search both efficient and accurate.

Implementation and Evaluation. Finally, we implement Weatherman and evaluate its accuracy on 117 smart meters and show that it localizes coarse (hour-level) energy consumption, wind, and solar data to within 16.68km, 9.84km, and 5.12km regions, respectively, on average. This represents significantly higher accuracy than recent work on solar localization [9], which i) only localizes solar energy data based on its solar signature, and not its weather signature, and ii) requires fine-grained second- or minute-level data and is not accurate using coarse hourly or daily data. We also evaluate how accuracy varies based on how well energy consumption data varies with outdoor temperature, which is a function of multiple factors, including the local climate, characteristics of the building's HVAC system, and the tightness of the building's envelope.

II. Background

Weatherman assumes it is given anonymous energy data that includes only a time-series of energy readings at a coarse resolution, e.g., every hour, with no other metadata. Weatherman's goal is to then analyze this anonymous energy data to infer the location—a latitude and longitude—of the smart meter that collected it. Weatherman currently focuses narrowly on localizing "pure" energy data, e.g., from either consumption, wind, or solar, and not "net" meter data that combines two or more types. Localizing net meter data that combines two or more data sources is future work.

To localize energy data, Weatherman searches a database of historical weather data to find a location where the weather data best correlates with a *weather signature* extracted from the energy data. Constructing a massive historical weather database from public sources, such as Weather Underground, that includes thousands of locations is not challenging. Our current weather database stores temperature, wind speed, and cloud cover each hour for 35,000 locations in the U.S., but could be expanded to other areas. We discuss more details of our prototype's weather database in §IV. For each type of energy data, Weatherman leverages a different physical model based on how that energy data relates to the location's weather to extract a weather signature. Below, we describe the physical models Weatherman uses to generate a weather signature for energy consumption, wind, and solar data.

A. Energy Consumption-Temperature Model

The dominant fraction of energy consumption in residential homes is due to space heating and cooling, which accounts for over 48% of energy usage [10]. The energy consumed for heating and cooling generally correlates with the outdoor temperature. This relationship is captured by the *degree-day* metric (in units of degree-time), which is the integral of the degrees above or below a specified base temperature over time for cooling and heating, respectively [11]. The base temperature represents the "balance" point at which no cooling or heating is required, and is typically estimated as 18C (or 65F) for buildings. The energy required to heat or cool a building is then modeled as being directly proportional to the number of heating or cooling degree-days, respectively. To illustrate this relationship, Figure 1 plots a home's daily energy usage on the y-axis, and the daily degree-days on the x-axis, over summer. We use a base temperature of 18C, so a degree-day less than 0 is a day where the temperature was always less than 18C.

The degree-days metric linearly correlates with energy consumption, with higher slopes indicating a greater correlation between changes in energy consumption and changes in outdoor temperature. However, the slope of the line and base temperature(s) vary significantly across buildings based on multiple factors, including the local climate, type of HVAC system, building insulation, and user behavior. For example, homes in San Diego, California, which has a mild local climate with a constant temperature near 65F, may have a low correlation, since the temperature is steady and HVAC is often not required. Of course, homes must have electric HVAC to exhibit a large slope. Since all cooling is electric, slopes are typically higher during warmer months. In contrast, only $\sim$38% of U.S. homes use electric heating, so 62% of homes will have lower slopes during the winter

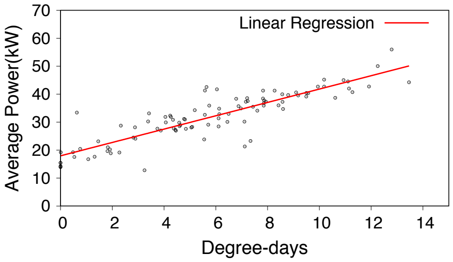


Figure 1. Daily average power and degree-days over the summer for a home with central air conditioning.

months. Even in locations that exhibit large changes in temperature, a building's insulation will affect the slope: a less insulated building will have a higher slope, since it will require more energy to maintain the base temperature when the outdoor temperature significantly deviates from it. Finally, the base temperature is also a function of user behavior, as it depends on the thermostat setpoint. Given these factors, the localization accuracy of energy usage data varies widely across buildings. As we discuss, our evaluation focuses on summer months where the correlation between energy usage and temperature is likely to be higher.

B. *Wind Energy-Speed Model*

The relationship between wind speed and wind energy generation is much simpler, since 100% of wind energy is a function of wind speed. Wind power generation is based on the cubic function below, where A derives from the turbine's rotor area, ρ is the air density, and v is the wind speed.

$$P = \frac{1}{2} A \rho v^3 \tag{1}$$

Wind turbine designs also dictate cut-in, rated, and cut-out thresholds that represent the wind speed at which power generation starts to increase, stops increasing, and terminates, respectively. At low wind speeds under the cut-in speed, there is not enough power to overcome the friction of the rotor. After the cut-in wind speed, power then increases cubically up to the turbine's rated wind speed, where its generator limits power to a constant output. The turbine generates this constant power up to a cut-out wind speed that can damage it, at which point the turbine engages brakes and power output drops to zero. While these thresholds vary based on a wind turbine's size and design, typical cut-in wind speeds are 3-4 meters per second (m/s), rated speeds are 12-17 m/s, and cut-out speeds are ∼25 m/s. Figure 2 is a scatterplot of hour-level wind generation and speed measurements with annotations of the turbine's cut-in speed, cubic function, rated speed, and cut-out speed.

C. *Solar Energy-Cloud Cover Model*

Solar energy embeds perhaps the most precise location information. Figure 3 shows the solar output of a small rooftop solar site as a function of the measured Global Horizontal Irradiance (GHI) in W/m^2 using a pyranometer at the same location. As expected, the relationship is almost perfectly linear, since solar modules translate irradiance directly into power with some efficiency loss. The small imprecision in the graph is due to minor temperature effects, which cause efficiency to decrease as the temperature rises.

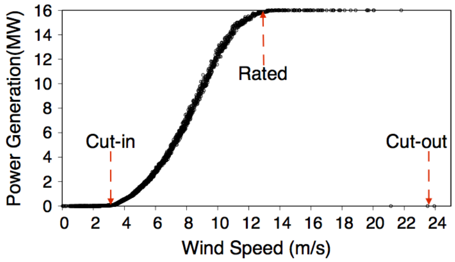


Figure 2. Hourly measurements of wind power and speed.

Unfortunately, unlike temperature and wind speed, most weather stations do not include pyranometers, and thus do not report ground-level irradiance. Thus, to localize solar data, we cannot simply correlate ground-level irradiance measurements with solar output. As a result, we use the coarse sky condition information reported by weather stations, which is typically measured in *oktas* that represent how many eighths of the sky are covered in clouds. Oktas range from 0 (completely clear sky) to 8 (completely overcast). The sky conditions reported by the National Weather Service translate directly to oktas [12] where "Clear/Sunny" is <1 okta, "Mostly Clear/Mostly Sunny" is 1-3 oktas, "Partly Cloudy/Partly Sunny" is 3-5 oktas, "Mostly Cloudy" is 5-7 oktas, and "Cloudy" is 8 oktas.

III. Weatherman Design

Weatherman uses the physical relationships above to search a large weather database to determine the location with weather that best correlates with a weather signature extracted from the energy data at each possible location. Since Weatherman assumes energy data is anonymous, it makes minimal assumptions about the associated metadata. For example, Weatherman does not assume the type of energy data is given, since classifying the data as energy consumption, wind, or solar is straightforward. Weatherman also does not require the associated units of energy data, e.g., watt-hour or kilowatt-hours, or their sign, as the correlation of energy with weather does not depend on the magnitude. In addition, Weatherman also supports different assumptions about the metadata information encoded in the timestamp. The specificity of the timestamp simply increases or decreases the Weatherman's search space. By default, Weatherman assumes the timestamp includes the date and hour. However, if the timestamp does not include the date, Weatherman simply correlates weather signatures for every possible daily time offset at each location, which increases the search space by 365× over a year.

A. *Weather-based Localization Challenges*

A naïve approach to weather-based localization is to directly correlate the time-series of energy data with the

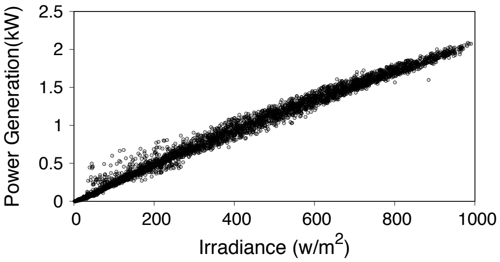


Figure 3. Solar energy is a linear function of solar irradiance.

time-series of a particular weather metric at each location in our weather database. There are many functions that quantify how well two time-series correlate, enabling a ranking of locations based on how well energy data matches a weather metric. For example, the Pearson Correlation Coefficient (PCC) is a measure of the linear correlation between two variables, computed as the covariance between the variables divided by the product of their standard deviation.

A naïve approach simply selects the locations with the highest PCC. Unfortunately, this naïve approach is imprecise, as energy data, itself, does not highly correlate with weather. For example, while changing weather instantly affects wind and solar energy, there is often a lag in the effect on energy consumption as a building heats up or cools down, which simple "instantaneous" correlation coefficients, such as the PCC, do not capture. Thus, Weatherman extracts a custom signature for each location that accounts for such effects in each type of energy data. In addition, searching all locations in a large weather database can be highly inefficient, since Weatherman must extract and compare a weather signature from energy data based on each location's weather metrics over a long period of time, e.g., multiple months to years. As a result, to improve efficiency, Weatherman first extracts weather signatures on coarse day-level data to filter the possible locations, as each type of energy also correlates with weather each day.

B. General Weather-based Localization Approach

Weatherman uses the same general approach to localize energy consumption, wind, and solar data. Weatherman extracts a custom weather signature from the energy data at each location in its weather database. To improve efficiency, it first extracts and correlates this weather signature at each location with coarse day-level weather data. Using day-level data both reduces the size of the input by 24×, and thus increases efficiency, and, in the case of energy consumption, also mitigates the impact of a variable lag in the energy response to temperature changes (since this lag is only evident at hour-level). Weatherman uses the day-level analysis to first filter possible locations by clustering the points using k-Means clustering, and then selecting the cluster with the highest average correlation. Filtering is important in reducing the large search space of locations.

Weatherman then re-computes the correlation using higher resolution data, and finds the weighted geographic midpoint of locations in the cluster (based on the magnitude of the correlation with each location) to estimate a final location. As we discuss below, the only differences between energy consumption, wind, and solar data is the method of extracting the weather signatures, the weather metric used for correlation, and the specific correlation function.

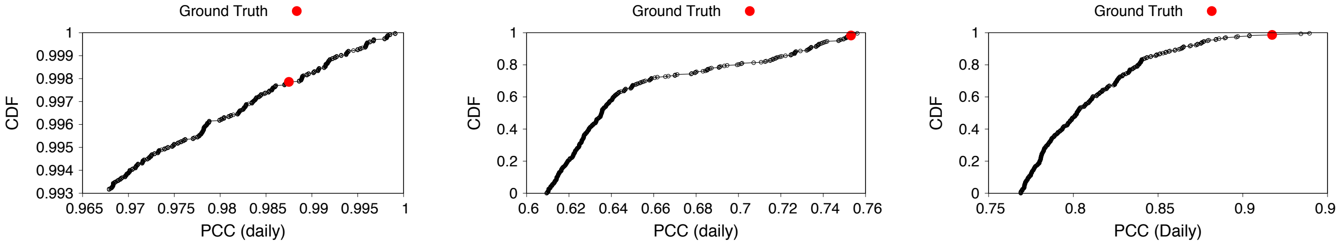
C. Energy Consumption Weather Signatures

Based on the degree-days model from §II, when correlating with each location, Weatherman removes energy consumption datapoints whenever the corresponding temperature is below the typical 18C base temperature. We assume energy consumption is linear with degree-days above a base temperature, and simply compute the correlation using the PCC between the daily energy and temperature data. Note that the daily correlation is robust to changes in user behavior, which are most prevalent within a day, e.g., from setting a programmable thermostat schedule that differs over the day, rather than across days. Figure 4(a) shows the CDF of the PCC for all locations, with the ground truth indicated as a red dot. This graph is for the same home as in Figure 1, and filters the locations from 35k to $\sim$300.

Unfortunately, for higher resolution hourly energy data, there is typically a variable lag between the increase in temperature and the corresponding increase in energy consumption, as it takes time for a building to heat up and for its thermostat to detect this and activate the HVAC system. This lag is variable, as it depends on the thermostat setting, which may vary, and the tightness of the building's envelope. As a result, the impact of a temperature increase is often not observed in energy data for an hour or more. Thus, the PCC does not work well with hour-level data, since it only considers the correlation between each two points in time.

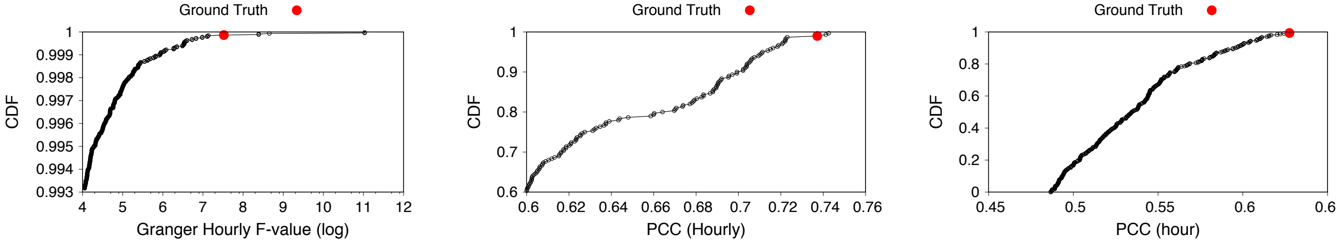
In this case, Weatherman applies Granger causality analysis [13], which captures the extent to which changes in one variable predict (or lag) another over time using an F-test. Note that, unlike the PCC, Granger causality analysis does not require that changes be linearly correlated, only that they lag and have the same direction. Computing Granger causality is more computationally-intensive than computing the PCC, since it searches over multiple possible lag values. As a result, performing Granger causality at hour-level over 35k locations is time-consuming. For example, a full search, assuming the date and hour are well-known, takes 8.5 hours using 80 high-end data center servers. If the date is not included in the timestamp, the search would take $\sim$8.5 $*$ 120=1020 hours (42.5 days) on the same set of servers, since we only conduct this search over the summer months. Thus, we only performGranger causality analysis over the filtered list of sites using the daily data analysis above, which takes $\sim$5 minutes.

Figure 5(a) shows the CDF of the Granger causality of the hourly energy data (using an F-test with a p-value$<$0.001). In this case, the final weighted geographic midpoint of these



(a) Energy consumption (b) Wind energy (c) Solar energy

Figure 4. CDF of correlation analysis across all locations for daily energy consumption, wind energy, and solar energy data.



(a) Energy consumption (b) Wind energy (c) Solar energy

Figure 5. CDF of correlation analysis across all locations for hourly energy consumption, wind energy, and solar energy data.

locations results in an estimated location 6.14km from the actual location (and within the same town). Note that the home is 4.1km from the nearest weather station, which has the fifth highest correlation in this case.

D. Wind Energy Weather Signatures

As Figure 2 shows, the relationship between wind power and speed is defined by a piecewise function based on the cut-in, rated, and cut-out speeds. A simple approach for extracting a wind weather signature would be to focus on just one part of this function. However, this would remove useful information. Instead, Weatherman projects this piecewise function onto the single line $y = 0$ (where y is the energy generation and x is the wind speed), such that the wind power data as a function of wind speed after this projection should be zero at the correct location. Since PCC and other correlation coefficients are undefined when the variance of one variable is zero, we rank locations based on their average absolute value after the projection, i.e., the average perpendicular distance from $y = 0$.

Weatherman does not alter the energy datapoints that correspond to wind speeds from 0-3m/s and >22m/s, since these should already map to zero. For datapoints in the range 4-13m/s, we first take the cube root of the energy data, perform a linear regression, and then find the distance each datapoint is from this line. We then project these points by replacing their original energy generation value on the y-axis with this distance value on the y-axis. For datapoints in the range 14-21m/s, we find the horizontal line that minimizes the root mean squared error with the datapoints, and then subtract the y-value of this horizontal line from the y-value of each datapoint. After this projection, wind power data that perfectly correlates with wind speed will lie at y=0.

We perform the same projection when filtering based on daily and hourly data, as described above. After performing this projection, Weatherman proceeds based on the basic approach above, where the weather metric is wind speed and the correlation function is the average of the absolute value of the projected data. Figure 4(b) and Figure 5(b) show the CDF of this average across all locations for the daily and hourly data for the data in Figure 2, with the ground truth location indicated by the red dot. We again are able to filter from 35k to ~300 sites using the daily energy data. Here, the nearest weather station to the actual location ranks fifth and the geographic midpoint of the selected locations is 24.37km from the target site. We then perform the same analysis on the filtered sites using the hourly data. In this case, the nearest weather station ranks third, and the geographic midpoint of the filtered locations is 3.87km from the location.

E. Solar Energy Weather Signatures

Solar power has a near linear correlation with solar irradiance, which is largely determined by cloud cover that is measured by weather stations in oktas, as discussed in §II. Unfortunately, raw solar generation does not directly correlate with oktas, as solar output varies over both the time-of-day and the day-of-year. Since these variations are a function of location, Weatherman does not know them precisely. However, we can roughly estimate the maximum generation potential P_s of solar by observing that the average clear sky irradiance, which is a well-known function of time at each location based on a site's efficiency, tilt, orientation, etc. should be an upper-bound on solar output, as described by the equation below. Here, β is the tilt angle, ϕ is the orientation angle, Θ is the Sun's zenith angle, α is the Sun's azimuth angle, $I_{incident}$ is the clear sky solar irradiance, and k is a module-specific parameter that combines a module's efficiency and size. The solar angles and clear sky irradiance are themselves a function of latitude, longitude, and time [14]. We can search for the parameters, e.g., latitude, longitude, efficiency, tilt, and orientation, that defines a valid solar curve that yields the tightest upper bound.

$$P_s(t) = I_{incident}(t) * k * [\cos(90 - \Theta) * \sin(\beta) * \cos(\phi - \alpha) + \sin(90 - \Theta) * \cos(\beta)] \quad (2)$$

We search for the parameters above as described in prior work [9], [15], [16]. Specifically, we first use prior work on localization using solar signatures to estimate a location by associating the first, last, and maximum hour of generation with the time of sunrise, sunset, and solar noon [9]. Note that we use this search only to provide a rough estimate of the hourly maximum generation; the latitude and longitude we find are not accurate for localization. Accurate localization based on the solar signature using hourly data is not possible, as it has a maximum accuracy of at most 1656km based on the speed of the Earth's rotation. Given this rough location, we then conduct a binary search for the efficiency, tilt, and orientation parameters that represents the tightest upper-bound on the data, as described in prior work [15], [16], since solar generation is bounded by the maximum clear sky irradiance at any time.

The model above does not account for temperature, which increases solar efficiency by some percentage c for every degree Celsius decrease in temperature. Thus, when extracting the weather signature for each location, we use the same model from prior work to adjust for these temperature effects [15], [16]. This approach conducts a binary search for the temperature coefficient c that results in the tightest upper bound on the data. That is, the approach adjusts the efficiency parameter k above at each time t for each c based on the equation below, where $T_{baseline}$ is the temperature at the time where the model above is closest to the actual solar data and T_{air} is the temperature at every time t.

$$k'(t) = k * (1 + c * (T_{baseline} - T_{air}(t))) \quad (3)$$

Since temperatures are different at each location, we must repeat this process at each location. This search provides a temperature-adjusted estimate of the maximum solar generation for each hour and day at each location. Weatherman then normalizes the daily and hourly data in its weather signature at each time period by dividing each data point by this maximum estimated solar generation. This normalized solar output (relative to the maximum possible clear sky output) should linearly correlate with the cloud condition in oktas reported by weather stations. As a result, Weatherman directly uses the PCC to quantify this correlation.

Figure 4(c) and Figure 5(c) show the CDF of the PCC across all locations for the daily and hourly solar data. We again filter from 35k locations to $\sim$300 locations using the daily data, and then perform analytics using the hourly data. The nearest weather station ranks fourth in the daily data with the geographic midpoint of the filtered locations 13.53km from the actual location. The nearest weather station ranks second using the hourly data with the geographic midpoint an estimated location 2.05km from the actual location. Again, the nearest weather station is 12.73km away from the actual location, so Weatherman's estimate is closer than any single point.

IV. Implementation

We implement Weatherman in Python, and make it and the data for this paper publicly available at the UMass Trace Repository.[1] We use the scikit package, which functions for PCC and Granger causality analysis. We implement a standard approach for finding the weighted geographic midpoint. Finally, we use the Pysolar Python package for estimating the clear sky irradiance [17]. Note that we set the thresholds for filtering the daily and hourly data based on an empirical analysis. We experimented with different thresholds in this range, and it did not significantly change the results. For the daily data, we use k-Means clustering with k equal to 50 to generate at least as many clusters as there are states. In general, larger values of k do not significantly affect the results. We then re-compute the correlation function of the locations in the selected cluster using the hourly data, and then compute their weighted geographic midpoint. We build our weather database by fetching data from DarkSky's weather data API. [2] Our database currently stores hourly and daily temperature, wind, and sky condition data from 35,000 weather stations in the U.S. The database includes data over four-month period from June to September, 2016.

V. Experimental Evaluation

§III illustrates Weatherman's approach on an example building, wind site, and solar site. In this section, we evaluate accuracy across many sites, and highlight how it varies across sites with different characteristics. Our evaluation uses energy consumption data from a sample of 100 homes in the Pecan St. dataset [18], as well as 10 solar sites and 7 wind sites. Each home in the Pecan St. dataset is located in the Pecan St. neighborhood near Austin, Texas.

A. Energy Consumption

Figure 6 shows Weatherman's localization accuracy for the Pecan St. homes, in terms of the distance between the location Weatherman infers and the ground truth location of the home. We sort homes by the slope of their average energy usage versus degree-day line, as depicted in Figure 1, which appears as a number above each bar. We also color each bar based on whether the home has zero, one, or two air conditioner circuits sub-metered, as we expect homes with more air conditioners to have a higher degree-day slope. We also experiment with two different timestamp meta-data assumptions: one where we know each point's date and hour

[1] http://traces.cs.umass.edu/

[2] http://darksky.net

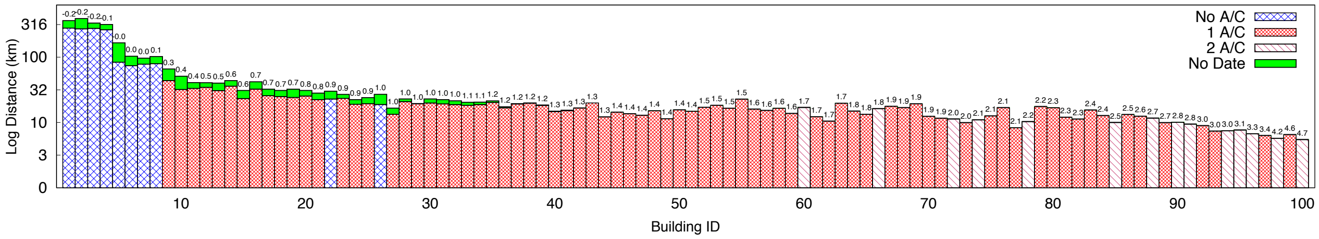


Figure 6. Weatherman localization accuracy for 100 different homes in Texas (on a log-scale). Homes are sorted by their degree-day slope, which appears atop each bar. The bar color indicates whether the home has zero, one, or two air conditioners.

(but not the timezone), and one where we only know the hour (but not the date or timezone). We indicate the decrease in accuracy from removing the date by placing an additional green bar atop each bar. Note that the y-axis has a log scale.

The graph shows that homes without air conditioners have a relatively flat degree-day slope, which indicates that their energy consumption does not change significantly with outdoor temperature. The first four bars of the graph have a localization accuracy of >200km and exhibit a *negative* degree-day slope, such that their energy consumption *decreases* as the temperature increases. Four other homes with no air conditioners have a non-negative slope in the range 0.0-0.1 and thus have a slightly better localization accuracy of ∼80km. There are two other homes without air conditioners that exhibit much higher degree-day slopes 0.9-1.0, and thus yield better accuracy of ∼20km. These homes likely operate other temperature-dependent loads.

As expected, homes with a single air conditioner exhibit degree-day slopes ranging from 0.3-4.6, and exhibit much higher localization accuracy, ranging from 5-40km with an average accuracy of 16.98km. Homes with two air conditioners tend to have an even larger degree-day slope and thus a higher average localization accuracy of 11.89km on average. Here, accuracy is a roughly linear function of degree-day slope, so homes with lower degree-day slopes, whether due to their local climate, HVAC system, building insulation, or user behavior, have a lower localization accuracy. We also observe that removing the timestamp's date does not significantly alter the localization accuracy: for homes with degree-day slopes >1.2 (near Building ID 37) it does not change, and for homes with degree slopes <1.2 it only slightly decreases. This shows that weather signatures are distinct, not only across locations, but also across time.

B. Wind Energy

Figure 7 shows the localization accuracy for 7 wind sites in Washington (#1), Idaho (#2), California (#3), Colorado (#4,6), Wisconsin (#5), and Texas (#7). In this case, we sort the sites by the variance in the average wind speed at their location over a year. We use variance as a proxy for the uniqueness of a location's weather signature, since the more the wind speed varies, the more opportunity Weatherman has to distinguish one location from another. As expected, the localization accuracy increases as the variance in wind speed at a location increases. In this case, the highest variance yields the highest localization accuracy of ∼3km, while the lowest variance yields the lowest accuracy ∼21km. Thus, wind localization is slightly more accurate than energy usage localization (for homes with air conditioners). For wind energy, removing the date from the timestamp also has little effect on the accuracy (indicated by the green bars as before).

C. Solar Energy

Figure 8 shows Weatherman's localization accuracy for 10 solar sites, as well as the accuracy for prior work on SunSpot [9], which localizes using a site's solar signature. The solar sites are in North Carolina (#1), Washington (#2), Colorado (#3-5), Texas (#6), Wisconsin (#7), Massachusetts (#8,10), and Ohio (#9). In this case, for SunSpot, we localize using minute-level data, while for Weatherman we localize using hour-level data. Similar to above, we sort the sites by the variance in their location's sky condition data, which is listed atop each bar. Using the same intuition as for wind, the more variable the sky condition, the more opportunity Weatherman has to distinguish one location from another. As above, we see that Weatherman's accuracy improves as the variance increases, with the most variable site having an accuracy of ∼2km. In addition, we see that solar localization accuracy based on weather is typically higher than either energy consumption or wind with all the sites having an accuracy between 2-7km. In general, the relationship between solar power and cloud cover (and temperature) is more direct than the similar relationships with energy consumption and wind, since solar is a purely electric device, while the other two involve more complex mechanical relationships.

We also see that weather-based solar localization is significantly more accurate using hour-level data than SunSpot using minute-level data. In particular, the *worst site* for Weatherman has an accuracy of 6.86km, while the *best site* for SunSpot has an accuracy of ∼12km. In addition, SunSpot has more variable accuracy, indicating that its solar signature is less robust than Weatherman's weather signature. Finally, we again see that removing the date from the timestamp has a minimal effect on localization accuracy.

VI. Related Work

As noted above, the work most similar to Weatherman is SunSpot [9], which localizes pure solar data using a

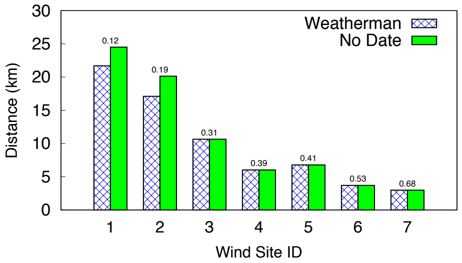


Figure 7. Localization accuracy for 7 wind sites, sorted by variance in wind speed, which appears atop each bar.

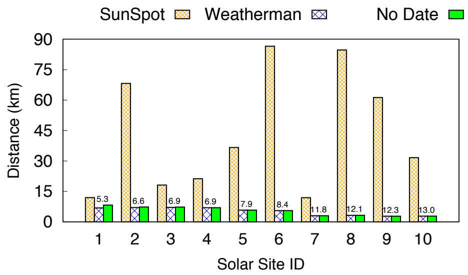


Figure 8. Localization accuracy for 10 solar sites, sorted by variance in sky condition, which appears atop each bar.

solar signature, specifically by inferring the time of sunrise, sunset, and solar noon. Weatherman shows that weather-based localization is *significantly more accurate* using *much lower resolution data* and requiring *much less data*. In particular, Weatherman's average solar localization accuracy using hour-level data is 5.12km, which is more accurate than SunSpot's accuracy using data that is 60-3600× lower resolution. In addition, SunSpot requires more than six months of data, since it needs data in both the spring/summer and fall/winter to pinpoint an accurate latitude, while our evaluation here only used data from four months in the summer. Finally, Weatherman is more general and also capable of localizing energy consumption and wind energy data to similar (or better) levels of accuracy.

There have been numerous papers on privacy-preserving techniques for energy data. These techniques generally focus on obscuring identifiable patterns in high resolution energy data, e.g., second-level or minute-level, using a controllable power source, such as a battery [6], [7] or a water heater [19]. These techniques are likely not effective in preventing weather-based localization, since it requires only coarse day- and hour-level data. In general, the battery and water heater capacity required to significantly alter day- and hour-level energy usage over a long period is prohibitively expensive. In addition, we also show that even modifying energy data to eliminating timestamp metadata, e.g., by not including the date or hour, does not significantly affect Weatherman's accuracy. Thus, preventing weather-based localization represents a challenging problem, which we plan to explore as part of future work.

VII. Conclusion

We present Weatherman, which leverages a suite of big data analytics techniques to localize anonymous energy usage, wind, and solar data. Weatherman shows how access to large archives of publicly-available, and seemingly innocuous, sensor data can introduce serious privacy threats. Our work shows that weather-based localization is highly accurate for multiple types of energy data. In particular, we show that Weatherman localizes coarse (one-hour resolution) energy consumption, wind, and solar data to within a radius distance of 16.68km, 9.84km, and 5.12km. These results are *significantly more accurate* using *much lower resolution* and *much less* energy data than prior work on energy-based localization, which only localized solar data to within ∼20km using second-level energy data.

Acknowledgements. This research is supported by NSF grants CNS-1645952, IIP-1534080, CNS-1405826, CNS-1253063, CNS-1505422, and the Massachusetts DOER.

References

[1] "Bidgely," http://bidgely.com, Accessed June 2017.

[2] "Onzo," http://www.onzo.com/, Accessed June 2017.

[3] "Sense," https://sense.com/, Accessed June 2017.

[4] A. Molina-Markham, P. Shenoy, K. Fu, E. Cecchet, and D. Irwin, "Private Memoirs of a Smart Meter," in *BuildSys*, November 2010.

[5] "Voluntary Code of Conduct (VCC)," U.S. Department of Energy, Tech. Rep., January 12 2015.

[6] S. McLaughlin, P. McDaniel, and W. Aiello, "Protecting Consumer Privacy from Electric Load Monitoring," in *CCS*, October 2011.

[7] W. Yang, N. Li, Y. Qi, W. Qardaji, S. McLaughlin, and P. McDaniel, "Minimizing Private Data Disclosures in the Smart Grid," in *CCS*, October 2012.

[8] J. Malof, R. Hou, L. Collins, K. Bradbury, and R. Newell, "Automatic Solar Photovoltaic Panel Detection in Satellite Imagery," in *ICRERA*, November 2015.

[9] D. Chen, S. Iyengar, D. Irwin, and P. Shenoy, "SunSpot: Exposing the Location of Anonymous Solar-powered Homes," in *BuildSys*, November 2016.

[10] "Energy.gov Heating and Cooling," https://energy.gov/public-services/homes/heating-cooling, Accessed June 2017.

[11] "BizEE degree Days: Weather Data for Energy Professionals," http://www.degreedays.net/, 2017.

[12] "Weather.gov Forecast Terms," http://www.weather.gov/bgm/forecast_terms, 2017.

[13] C. Granger, "Investigating Causal Relations by Econometric Models and Cross-Spectral Methods," *Econometrics*, vol. 37, no. 3, 1969.

[14] M. Blanco-Muriel, D. Alarcon-Padilla, T. Lopez-Moratalla, and M. Lara-Coira, "Computing the Solar Vector," *Solar Energy*, vol. 70, no. 5, pp. 431–441, 2001.

[15] D. Chen and D. Irwin, "SunDance: Black-box Behind-the-Meter Solar Disaggregation," in *e-Energy*, May 2017.

[16] D. Chen and D. Irwin, "Black-box Solar Performance Modeling: Comparing Physical, Machine Learning, and Hybrid Approaches," in *Greenmetrics*, June 2017.

[17] "PySolar," http://pysolar.org/.

[18] "Pecan St. Inc." http://www.pecanstreet.org/, Accessed 2017.

[19] D. Chen, D. Irwin, P. Shenoy, and J. Albrecht, "Combined Heat and Privacy: Preventing Occupancy Detection from Smart Meters," in *PerCom*, March 2014.