# Leveraging Big Data to Develop Next Generation Energy Efficiency Programs and Energy Regulations

*Celina Stratton, Energy Solutions, Long Beach, CA*
*Mike McGaraghan, Energy Solutions, Oakland, CA*
*Teddy Kisch, Energy Solutions, Oakland, CA*
*Arshak Zakarian, Energy Solutions, Oakland, CA*
*Peter Borocz, Energy Solutions, Oakland, CA*
*Eric Rubin, Energy Solutions, Oakland, CA*
*Carolyn Richter, Energy Solutions, Oakland, CA*
*Mary Anderson, Pacific Gas and Electric Company, San Francisco, CA*

## ABSTRACT

"Big data" – the concept of developing and analyzing increasingly large sets of data – is becoming common as the tools for data collection become more powerful. In the energy efficiency industry, through the application of custom "web crawling" software, it is now possible to cost-effectively collect massive amounts of data to support improved market analyses for efficiency programs and standards. In fact, many online retailers are already set up to facilitate this type of ongoing data collection through their websites' application programming interface (API) specifications, enabling the near real-time collection of thousands of data points covering price, efficiency, product size, and many other product performance attributes. This paper discusses potential uses for web crawler data both to develop more effective energy efficiency incentive programs and energy codes and standards, and to conduct evaluations of these efforts. First, we present five case studies of ongoing web crawler data collection efforts for selected consumer electronic, appliance and lighting products. Next, we discuss the potential for leveraging this data to conduct macro-analysis of market trends. Specifically, we show how big data can greatly improve the accuracy of key metrics, such as incremental measure costs and efficiency distributions for given products. In sum, big data may prove to be a game-changing tool for the energy efficiency industry to maximize energy savings for the next generation of energy efficiency initiatives.

## Introduction

The development of energy efficiency programs, as well as the evaluation of those programs, relies on calculations of energy savings and cost-effectiveness of the proposed measures. Ideally, these calculations would be calibrated to account for the dynamic nature of measure baselines, product pricing, incremental measure cost (IMC), product performance characteristics, measure savings, useful life, life cycle cost savings, and naturally occurring market adoption (NOMAD) rates. Most products experience significant market changes over time (whether rapid or gradual) in several or all of these categories, and both the analysis of past trends and the forecasting of future trends has been constrained by historical methods for collecting these data. These market shifts result from a number of factors including improved manufacturing processes and industry learning, availability and cost of raw materials, increased industry competition, or fluctuating market demand, and can render data from traditional collection methods inaccurate within a few months.

Fortunately, the tools for data collection are becoming increasingly powerful, and the development and analysis of large datasets, or use of *big data*, can greatly expand the types of analysis available. *Web crawler* software tools are showing significant promise by consolidating the collection of price points and multiple product attributes in real-time and gathering the data in increasingly large quantities, cost-

effectively. This method can dramatically improve efforts to quantify price-performance relationships and enable more reliable and defensible forecasts for product performance and pricing over time. Web crawler tools have several limitations, which will be discussed in this paper, but these tools can be useful for many different applications within the energy efficiency field to improve program design and evaluation practices, in order to increase energy savings impacts.

This paper briefly describes the limitations of recent market analysis efforts and an overview of custom web crawler software within the energy efficiency context. It then provides four examples of ongoing efforts that use web crawlers to improve understanding of price-performance relationships, regional product availability, and appliance standards compliance trends. It concludes by offering suggestions for how web crawlers and the data they collect can enable energy efficiency opportunities.

## Limitations of Recent Market Analysis Efforts

Historically, market analysis efforts have been constrained by data availability, analysis capacity, and costs. One example of a traditional market analysis strategy is conducting shelf surveys—market characterization efforts where researchers visit brick-and-mortar retailers and manually survey the products available. The surveys commonly collect pricing, product availability, and product performance data from a variety of retailers. Shelf surveys have been a great source of data for the industry; however, they are often limited due to the high cost of manual data collection and compilation, and the time required to complete. Further, they generally provide data relevant only to a snapshot in time of a particular market and comparison between similar studies at different points in time commonly allows for only tenuous trend analyses, since study design often varies.[1]

There are a variety of other data sources that track product pricing and market trends over time. For example, rebate programs themselves often collect valuable product price and sales information during program implementation, though in many cases the data are limited to the products that qualify for rebates, to the retailers participating in the program, or to the duration of time during which the program was running. Another data source is the US Bureau of Labor Statistics' Producer Price Indices (PPIs), which provide historical manufacturing costs for various appliances and equipment. The U.S. Department of Energy (DOE) recently began using these PPIs, in conjunction with cumulative shipment data from industry groups like the Association of Home Appliance Manufacturers, to develop experience curves for specific products. These product-specific functions showed the extent to which product manufacturing costs come down over time, as a function of the number of units manufactured. These experience curves allowed DOE to forecast future price reductions in its standards rulemakings, and to forecast a proportional decline in incremental measure costs (previously, DOE's standards analyses had always assumed constant IMC throughout the 30 year period of analysis) (Desroches et al. 2012). The implications of this change in DOE's analysis are significant—higher life cycle cost savings from proposed standards means DOE standards can be more aggressive—but the approach is still quite conservative, since higher efficiency products likely rely on newer technologies, which are expected to decline in price at a faster pace than the rate of decrease for the product class as a whole. In other words, pricing of higher-efficiency products is likely to decrease more quickly than the pricing of lower-efficiency products, but PPIs do not provide this level of granularity. While PPIs are readily available and relatively low cost, they generally do not take into account narrowly defined product sub-types or sizes, performance attributes (such as efficiency), geographic area, or sales channel, each of which could significantly affect price and IMC, depending on the product.

[1] There have been exceptions, including the Northwest Energy Efficiency Alliance's multi-year lighting tracking studies available at NEEA.org.

## Big Data

Developing and analyzing progressively larger sets of data is becoming common in the energy efficiency industry as the tools for data collection become more powerful through advances in information technology. The ability to collect wide swaths of data on an ongoing basis and the versatility of customizable software offers enormous potential for improving the analysis used for energy standards and market research. For a given product, big data has the potential to provide a detailed, accurate assessment of the current market, and if maintained, can build a foundation to forecast trends over time. One such big data collection model, described below and used in each of the case studies presented in this paper, uses automated web crawler software tools to track real-time product performance, price and availability data that can inform standards development, program design, and evaluation.

### Custom Web Crawlers

Web crawlers are specialized software tools that are programmed to track specific product information on retailer websites. Many existing web crawler services such as CamelCamelCamel.com and PriceGrabber.com cater to consumers, tracking price trends for specific models. These existing tools, however, don't provide the exact level and precision of data that would be most valuable to a price and performance data analysis. Customized web crawlers can be designed to pull granular data needed for energy efficiency measure analysis and to do so at regular intervals. In some cases, online retailers provide Application Program Interfaces (APIs) to allow interested parties easier access to data from their websites without interfering with the main sites that serve typical customers. Essentially, these APIs expose underlying product databases to simplify the web crawling process. Initial tool development and ongoing maintenance costs are greatly reduced when an appropriate API is available. If an API is not provided, a web crawler can be programmed to include *screen scraping* capability to extract the appropriate data from the retail site.

At a minimum, a useful web crawler will collect the following product attributes for a large proportion of the market: retailer, brand, model number, price (including regular price and sale price), and usually many other product specifications, including energy use, size/capacity, warranty, ENERGY STAR® certification (and/or other energy efficiency tier or rating), and a wide range of other performance features specific to the equipment type. It may be desirable to attempt to collect data regionally. For example, with some online retailers, such as Home Depot, online prices are displayed based on the assumed zip code of the user browsing the website. Web crawlers can be programmed to search from any zip code, so it is possible to analyze prices and availability by region, which can help identify influences of efficiency programs and standards on retailer stocking practices.

There are also a number of limitations of web crawler tools that need to be considered. First and foremost, web crawlers do not generally collect sales data, so the performance and price data collected are not sales weighted (though some sites such as Amazon.com do indicate "top sellers" or other strategies to differentiate popular items, which could be similar to the assessment of "shelf space" for certain products often included in store surveys). Additionally, in some cases average prices offered online may differ slightly from average prices offered in brick and mortar stores. For example, some online retailers appear to offer a larger range of products, including many very high-priced products, while brick and mortar stores may place more emphasis on providing their customers with only the best prices, and may therefore limit their offering. For both of these reasons, the products offered for sale online, and the price associated with them, should not be considered representative of all products, prices or purchases. Further opportunity exists to map sales data (for example sales data collected through an incentive program) to web crawl data, which would enable data analysts to establish sales-weighted assessments of product performance, energy use, incremental costs, and/or estimates of NOMAD of higher efficiency products.

Knowledge of utility rebate programs that may be reflected in online pricing (and how these rebate offerings change over time) is also important to consider in the design of a web crawler tool. If the prices offered online for more efficient products have been lowered by an upstream utility rebate, for example, and the team analyzing the data is not aware of this, any price analysis of this data will be entirely off-base.

Further, not all data are available online – many product categories are simply not sold through the internet, and certain retailers do not have an online presence at all – so web crawler tools cannot be used to support all program efforts. Lastly, for some products, retailer websites may not specifically list product energy rating or other performance metrics that may be of interest from the energy efficiency measure perspective, such as power factor. In these cases it is often possible to link the product pricing data obtained by a web crawler with product performance data available through other online databases (such as the ENERGY STAR qualified product list or other similar industry resources). To some extent, this model-matching strategy can be programmed to happen automatically, although it can be a challenge to achieve some model matches given the inconsistency in how retailers list manufacturer model numbers.

## Case Studies: Big Data in Action

The case studies highlighted below provide examples of using custom web crawler tools to collect market data and use it to improve understanding of market trends. These examples aim to overcome the data availability, analysis capacity, and cost barriers that have historically limited the ability to analyze market trends and make robust forecasts. The first two examples focus on using web crawler data to better understand relationships between product attributes and price for a variety of products, highlighting results for air cleaners and LED lamps. The hedonic price modeling method has been used for all products and is explained in detail using air cleaners as an example product.

### Overview of Hedonic Price Modeling Methods and the Development of an IMC for Product Efficiency

Based on the data sourced from the web crawler, we use a stepwise regression analysis to develop a hedonic price model and estimate the IMC attributable to increased product efficiency. Product efficiency, and its associated IMC, can be defined either as a discrete variable (e.g., ENERGY STAR vs. non-ENERGY STAR) or continuous variable (e.g., percent more efficient than Federal Standard). We have completed hedonic price modeling to inform decision-making for a wide variety of program applications and product categories, including various appliance standards and voluntary incentive programs, for products including air cleaners, clothes dryers, freezers, sound bars, home-theater-in-a-box, computers, and several different types of LED lighting products.

### Price and Performance Relationships: Air Cleaners

As part of the Pacific Gas and Electric (PG&E) and Sacramento Municipal Utility District's Retail Plug-load Portfolio program,[2] a custom web crawler was developed and the resulting price/feature data were analyzed using the hedonic price modeling methods to estimate IMC for five product categories, including air cleaners. To collect large amounts of product attribute data, a web crawler pulled data from seven popular retail websites in a single day in early 2015. Through this process, over 500 unique air cleaner price points, each with over 100 distinct product attributes were initially collected and were subsequently distilled down to 33 key product attribute variables based on the following considerations: [3]

[2] The Retail Plug-Load Portfolio (RPP) is a mid-stream utility energy efficiency incentive program designed to influence retailer decisions to stock and sell more energy efficient home appliances and consumer electronics in targeted product categories.

[3] There are several benefits to selecting among only a deliberately chosen subset of the possible predictor variables. When seeking to identify the best hedonic price model, the more product attributes that we evaluate, the more likely we are to incorrectly identify

- **Data Prevalence:** Frequency in which variables had available product data (i.e. product attributes were provided by retailer). For a product to be included in a multiple regression analysis, every variable being modeled must have a recorded value. Consequently, including variables with very low data prevalence decreases the sample size, and thus decreases the model's statistical power.
- **Expert opinion:** We interviewed product experts to determine which attributes they believed had the greatest impact on price and if there was a significant incremental cost for achieving increased product efficiency.

As a next step, we further reduced the number of significant variables from 33 to 11 based on how strongly they were correlated with product price.[4] While this approach does not control for the effects of other product attributes, it provides a first approximation of whether a variable may have an impact on price. For example, as Figure 1 demonstrates, there is a significant difference in price for the ENERGY STAR attribute, while 'Removes Mold' does not appear to have a significant impact on average product price.

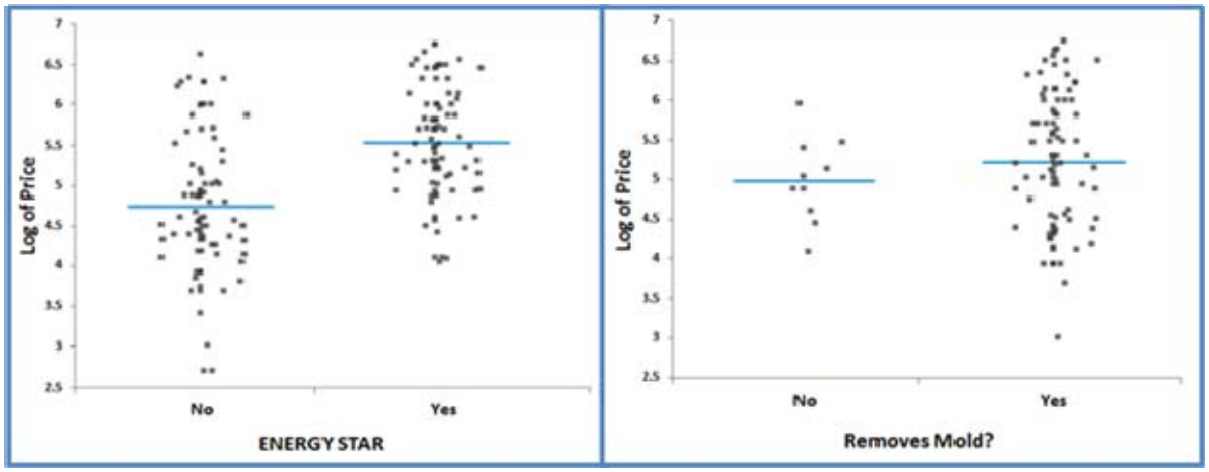


**Figure 1.** Box plot of Categorical Variables and their Effect on Price (blue line represents average price)

Based on these 11 product attribute variables, we split our sample size into two groups: a training dataset (representing 70% of the overall sample size) used to develop a set of best fit models that predict price and a test dataset (representing the remaining 30%) to test these models. We conducted a backwards stepwise regression on the training dataset to further refine which variables were most important in predicting price. The stepwise regression process removes variables one at a time that have multicollinearity problems[5] or are the least significant factors in predicting price until all variables are significant. Figure 2 provides an overview of the stepwise regression process and shows that brand was the first variable removed because it was found to be partially multicollinear with other product attributes (coverage and CADR), followed by several other attributes that were also found to have multicollinearity issues or to have the least significant unique contributions to predicting price. Based on these results, we developed three candidate models, ranging from three to five variables, to test on the remaining 30% of the dataset.

---

an attribute as being important when its true relationship with price is just noise. In addition, distilling the product attributes beforehand substantially decreases the degree of multicollinearity and missing data in the early stages of our backwards stepwise regression process, which eventually leads to a more realistic hedonic price model.

[4] To gauge how strongly product attributes influence price when analyzed in isolation, we plotted them against price and conducted simple linear regressions, one-way analysis of variance (ANOVA).

[5] When one or more product attributes are highly correlated (for example, freezer capacity and freezer weight), there is less unique variation in those attributes, and consequently their estimated effect on price are less precise. High multicollinearity may cause important attributes to appear insignificant and lead to unstable regression coefficients that are highly sensitive to changes in the data.

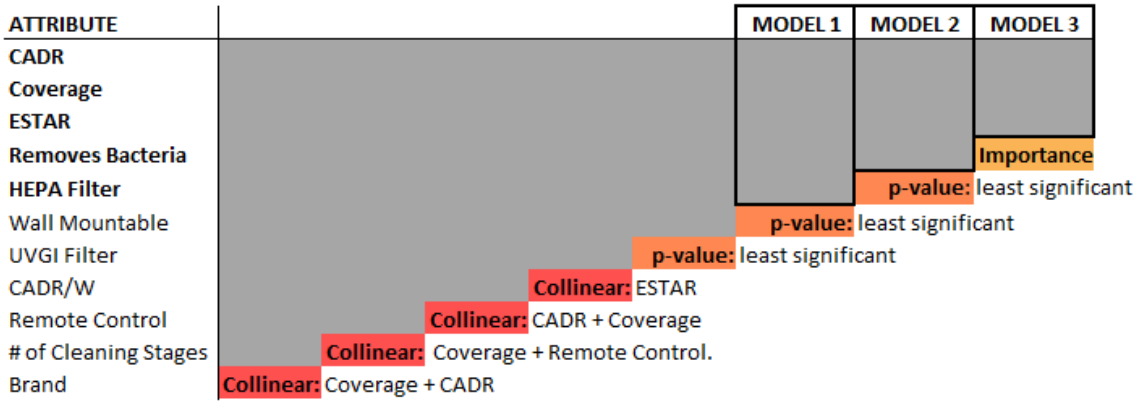


**Figure 2.** Stepwise Regression of Air Cleaner Variables – Demonstration of the Process of Eliminating Insignificant Variables

To select between the candidate models, we tested each model's ability to predict the actual prices of our test data set. Model 1 was the top-performing model and predicted prices with an average error of 3% (95% CI: -14% to 17%). This result suggests that each of the five attributes is important in predicting price. Based on this result, we re-fit this model to the entire dataset to determine the model's ability to predict price for the entire dataset (n=67). This model explains 81% of variation in price within the full dataset ($R^2$=0.81, adjusted $R^2$= 0.79).

**Determining the IMC of ENERGY STAR Certification**. Controlling for the other significant variables within the model, we identified that the impact of ENERGY STAR certification on price was an estimated increase of 56% of retail price on average, with a *p*-value of <0.0001. In other words, when controlling for other variables that impact price ('CADR', 'Coverage', 'Removes Bacteria', and 'HEPA Filter'), ENERGY STAR models are estimated to be 56% more expensive than non-ENERGY STAR models on average. As Figure 3 indicates, the difference in *average* price between ENERGY STAR Air Cleaners ($318) and non-ENERGY STAR Air Cleaners ($194) is $124. However, of this $124 difference, the hedonic model developed suggests that roughly $109 (88%) is attributable to ENERGY STAR certification, and this estimate is *statistically significant* (p<0.0001). This finding is consistent with information provided by our expert interviews, which identified that there appeared to be an IMC associated with ENERGY STAR products (likely due to a more efficient motor required to achieve greater efficiency). This value of $109 was therefore recommended as the final IMC to be used for this product category.

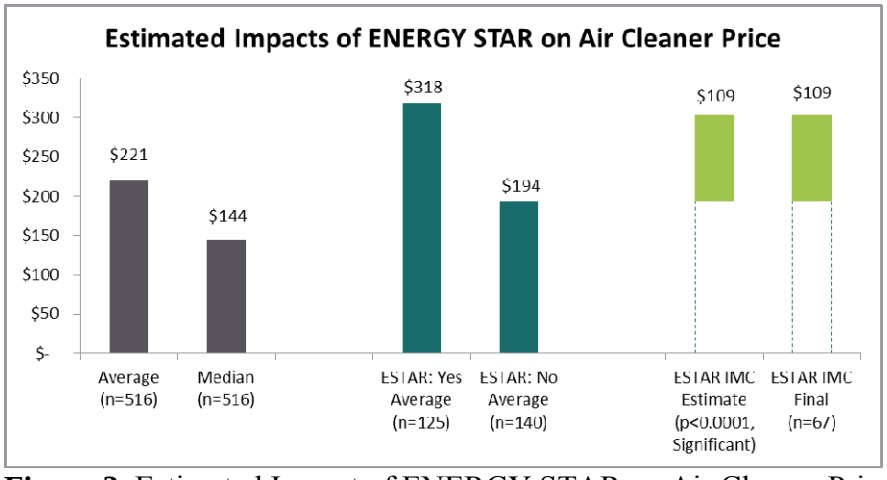


**Figure 3.** Estimated Impact of ENERGY STAR on Air Cleaner Price

## Price/Performance Relationships: LED Lamp Price Studies

Similar price and performance relationship analyses have been conducted utilizing web crawl data in support of other incentive programs and codes and standards (C&S) development efforts. For some products, like LED replacement lamps (i.e., LED light bulbs), data have been collected over time, either on a weekly or monthly basis, allowing for analyses of performance and price trends over time rather than the one-time snapshots provided by the regression analyses. This section presents work initiated in 2012 by the PG&E C&S team to analyze the relationship between performance and price for LED lamps, and how these relationships have been changing over time.

**Manual LED Data Collection**. In 2012, the PG&E C&S team began an effort to understand what aspects of LED lamp performance had the biggest impacts on end price, and the specific impacts associated with different aspects and levels of quality. At the time, we had not begun to use web crawler tools to collect lamp data, so in summer and fall of 2012, the PG&E C&S team undertook a time-intensive process to collect several hundred lamp price values from various online lamp vendors manually and correlate them to lamp performance characteristics. The team then conducted a multi-variable regression analysis of these lamp prices and characteristics, similar to the hedonic price modeling approach described above, to evaluate and refine a model that could help explain some product price fluctuations as a function of lamp performance (Young et al. 2014). This analysis was instrumental in understanding the interactions between price and various performance metrics and was used to inform the cost-effectiveness evaluations presented in a 2013 LED Codes and Standards Enhancement (CASE) Report. However, collecting the data manually was tedious and time consuming, and the data set was limited in terms of the variety of product types and performance features due to high collection costs. While useful for producing pricing for a single snapshot in time of the LED lamps market, a manual approach would not be cost-effective to collect this type of data on an ongoing basis.

**Web Crawler LED Data Collection and Results**. In fall 2013, the C&S team developed and rolled out a custom, retailer-based web crawler tool to automatically capture product pricing data for all LED products being sold at nine online retailers. Prices were initially gathered weekly and subsequently monthly and stored in a database that enables users to track product price fluctuations and trends over time. Price data are being collected for over 40 LED replacement lamp varieties, including both directional and non-directional replacement lamps, lamps with a wide variety of base types, and downlights/recessed retrofit kits, on an ongoing basis. Between November 2013 and April 2015, over 325,000 unique price points have been collected, corresponding to more than 7,500 unique LED product models offered from over 170 manufacturers.

In addition to real-time price and performance collection from retailer websites, the tool has been improved to link the data obtained from retailer sites to other publically available product performance data. Wherever possible, the tool maps the model numbers collected to those in the both the Lighting Facts Database[6] and the ENERGY STAR Qualified Products List[7] to obtain additional lamp performance information that can be linked to the price points collected online. Where a model match to Lighting Facts Database or ENERGY STAR is unsuccessful, the tool relies on performance information captured from the retailer. This linkage of product performance data allows users to view product price data as they relate to nuanced performance attributes not often advertised by retailers—including power factor, luminous intensity distribution patterns, or R9 (ability to accurately render red objects).

[6] http://www.lightingfacts.com/

[7] https://www.energystar.gov/products/certified-products

The following figures provide an example of the data being collected, and the types of trends and insights that can be drawn from it. One of the most notable trends is the high rate of price decline in the LED replacement lamp market. On average, online prices have dropped by about 25% in a little over a year of monitoring, while prices have declined more quickly for certain sub-groups of products. These figures highlight the unique trends observed within very specific sub-classes of products—for example we can see in Figure 4 that on a price per kilolumen basis, higher efficacy products are on average *less expensive* than lower efficacy products, a trend rarely observed in the world of energy efficiency measures. We can also see that the rate of price declines appears to be faster for the highest efficacy category: while per-kilolumen prices for lower efficacy products dropped by about 20%, average prices of very high efficacy lamps (>90 lumens per watt) dropped by about 40% over the same period.

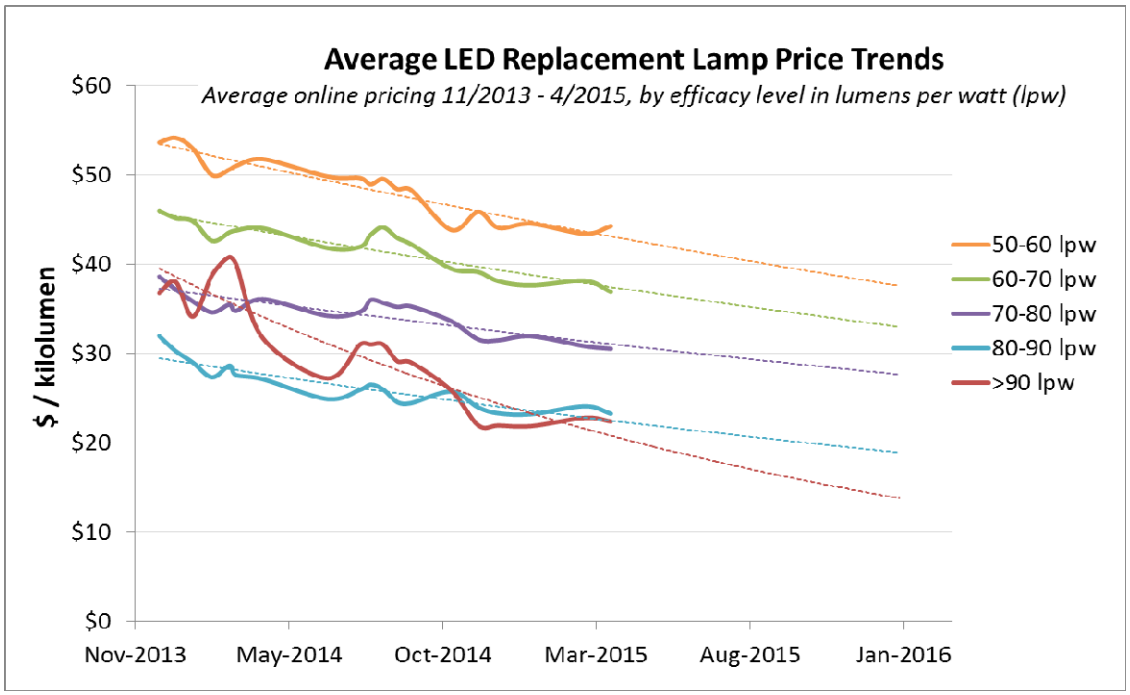


**Figure 4.** Average Online Pricing per Kilolumen for LED Replacement Lamps and Downlights, by Efficacy Level

Figure 5 demonstrates price trends for A-shaped lamps with different color rendering indices (CRI), a key metric utilized for qualification in California LED rebate programs and a new California residential building code requirement. The figure shows that A-lamps with CRI between 80-85 saw average per-kilolumen price reductions of about 35%, while high CRI (>90) average product pricing dropped by about 50%.[8] Understanding these types of trends can help program planners and standards developers to ensure that the performance targets they set stay ahead of technology improvement, and can help evaluators ensure that program designs correspond to demonstrated market trends.

[8] This analysis has removed the impact of any utility incentives to ensure a fair comparison of pre-rebate prices.

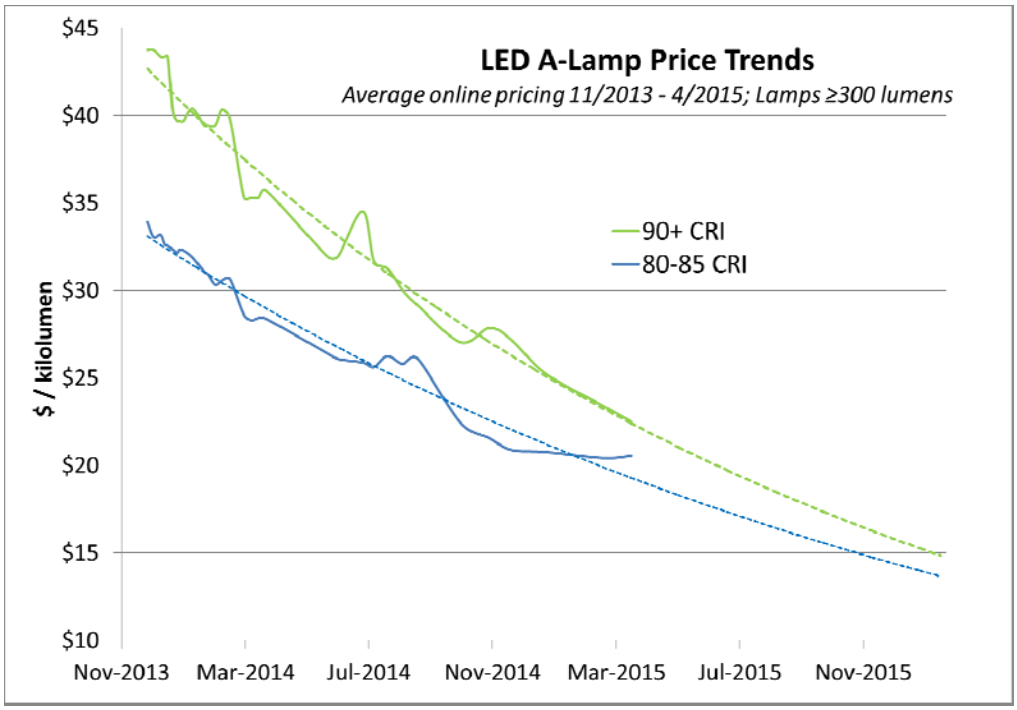


**Figure 5.** Average Online price per Kilolumen for LED A-lamps, by CRI Level

## Regional Product Availability Analysis

In addition to exploring price-performance relationships, we have done some basic analyses on the LED web crawler data set from an eight month period in 2014 to compare retail data from different cities. This analysis is relevant because of the regional nature of LED performance specifications. In 2013 the California Energy Commission (CEC) established a lighting quality specification for LED replacement lamps and required that LED products meet the new specification in order to be eligible for California investor-owned utility (IOU) rebates. The California LED specification established minimum performance requirements for a number of quality attributes, including color rendering (minimum of 90 CRI), color consistency, dimmability, rated life, warranty, and light distribution (Flamm et al. 2012). In 2013 California's building code (Title 24) also began to require LEDs installed in residential new construction to have a minimum CRI of 90 (if claiming credit as high efficacy lighting). We were interested to see if the building code and the California LED specification were having an impact on the availability of high performance LED products in California stores.

For this analysis, we chose to focus on 90 CRI because of its prominent role in both the voluntary rebate specification and the building code. For the analysis, we used web crawler data collected from online retail sites of two national retailers (Lowes.com and HomeDepot.com), with the web crawler programmed to collect data from zip codes in seven different cities.[9] Figure 6 indicates that on average LED lamps (and downlights) with CRI ≥90 were more widely available (more choices offered) in the four California cities compared to St. Louis, MO, Waltham, MA, or Houston, TX. In California cities, LED lamps with CRI ≥90

[9] Zip codes and corresponding cities included were: 02451 (Waltham, MA), 63139 (St. Louis, MO), 77087 (Houston, TX), 90014 (Los Angeles), 92122 (San Diego), 94102 (San Francisco), 94203 (Sacramento)

made up ~10% of the unique product price points collected from the online retailers; in the cities outside California lamps with CRI ≥90 made up ~8% of the unique product price points collected.[10]

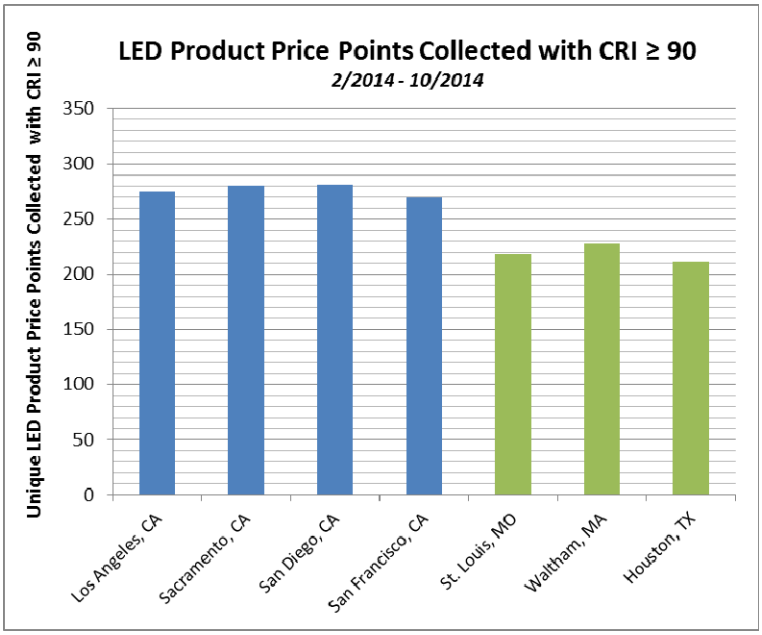


**Figure 6.** LED Product Price Points Collected with CRI ≥ 90 from two large retailers in select cities in the United States. Online price data collected 2/2014 – 10/2014

Though the difference in the number of 90CRI product offerings by region is not necessarily statistically significant, and there are other factors that could account for these differences, it seems plausible that the California LED Quality Specification and associated rebate programs, and CA's building standards, could be having an impact on the number of different 90+CRI LED products offered for sale in the State. Additional research into these methods would be helpful to expand upon their utility, but we believe analyses such as this could potentially be used by evaluators to help define the impacts of energy efficiency measures on the actual market they are intending to influence. Product availability and pricing can be compared in areas with and without the program being evaluated to better understand if the intended effects are being realized.

## Appliance Standards Compliance Analysis

In a separate analysis, the PG&E C&S Team is using custom web crawlers to understand appliance standards compliance trends. Web crawlers are currently collecting data from five online retailers for seven products in support of standards compliance analyses. One of these analyses aims to determine the duration of the lag time after a new standard takes effect, during which the existing stock of non-compliant products are sold or otherwise removed from inventory. Understanding these trends could be valuable to help determine when the evaluation of a standard should be conducted to most appropriately capture its effects. For example, several years ago, consultants working on behalf of the California Public Utilities Commission to conduct the Impact Evaluation Report for the Statewide Codes and Standards Program (Program Years 2010-2012), conducted two shelf surveys to assess compliance rates with the Federal standards for

[10] During each data collection event (17 total), the web crawler collected data on all the LED products offered for sale on the websites from each location. Thus, if a single lamp model was offered for sale from the same retailer location during each data collection event it would represent 17 unique product price points in the final dataset.

incandescent reflector lamps. These surveys were conducted one and two months, respectively, after the effective date of the new standards. The compliance rates observed in the shelf surveys were extremely low (~7%). However, low compliance immediately following the effective date of a standard is not necessarily indicative of perpetually low compliance rates, so that may not have been the appropriate time to conduct the analysis. Presumably, it takes several months or more for the existing stock of products *manufactured* before a standard effective date to run out, but no data is available on how long this period might be for different product types.

For the analysis presented here, we focused on clothes washers because new federal standards recently became effective–clothes washers *manufactured* on or after March 7, 2015, must meet new energy conservation standards. The analysis includes data collection prior to and after the standards came into effect. We then conduct a model matching analysis of product model numbers collected online with those in the CEC Appliance Efficiency Database (CEC Database) to determine if the product offered for sale is listed in the database.[11] In addition to model number matching between retailer sites and compliance databases, a secondary compliance check could be implemented by collecting energy performance metrics from retailers and comparing to the relevant standards requirements. However, in the case of clothes washers, the metrics utilized in the recent standards (Integrated Modified Energy Factor or IMEF, and Integrated Water Factor or IWF), are not yet commonly reported by retailers. As such, our team was not able to complete this secondary check.

The table below demonstrates the prevalence of clothes washer models being offered for sale before the standards effective date and in the two months immediately following the effective date that could not be matched to a model number in the CEC Database for the new standard. The table shows that the number of products not matched to the CEC Database is significant, but declining, as would be expected (older non-compliant products are expected to drop off as products run out of stock, and all new product offerings are expected to be compliant). Note that the products not found to be model matches to the CEC Database are not necessarily "non-compliant" products; many of these may have been manufactured before March 7 and are thus can still be legally sold until their stock runs out. Additionally, some of these products may in fact meet the new energy efficiency requirements, but have not been certified to the CEC Database yet.

**Table 1.** Analysis of Clothes Washer Models Offered for Sale Online Before and After the Effective Date of New Standards

| Data Collection Date | | Quantity of clothes washer models collected from 5 online retailers | Number of models that could not be matched to a model # in the CEC's Compliance Database for the new standards | Non-match rate |
|---|---|---|---|---|
| March 2015 | (before effective date) | 427 | 265 | 62.1% |
| April 2015 | (after effective date) | 408 | 251 | 61.5% |
| May 2015 | | 418 | 241 | 57.7% |

Conducting this analysis for additional products and for a longer duration (ideally beginning up to six months before an effective date and continuing for a year or more afterwards), would likely help shed light on the rate at which different markets react to effective dates for new standards. The results of these analyses and other similar analyses currently underway are intended to inform state and investor-owned utility efforts to understand energy savings potential from appliances efficiency standards. Surveying and collecting market information simultaneously with recent energy efficiency data can also help to identify areas of low compliance, where state and electric utility efforts can focus on increasing compliance.

[11] Certification to the CEC's Appliance Efficiency Database is a required part of compliance with California's Title 20 Standards.

## Opportunities and Conclusions

Big data has great potential for improving energy standards and energy efficiency program development. As discussed, recent market analysis efforts have been constrained by data availability, analysis capacity and costs, but case studies have demonstrated how web crawler tools can be used to develop more reliable methods for understanding (and forecasting) price-performance relationships, baseline energy use and measure savings potential, product availability trends, and standards compliance trends.

The case studies presented here represent initial explorations into the potential value of web crawler tools for developing large data sets to support program design and evaluation, though other advanced analyses may be logical next steps. For example, there is potential to increase the utility of web crawl data by combining it with sales data (such as data collected by utilities during rebate program implementation), to establish sales-weighted assessments of product performance, energy use, incremental costs, and/or estimates of NOMAD of higher efficiency products. Another opportunity would be to develop relationships between web crawler data and shelf surveys to help validate both data sources. Since online pricing does not necessarily correspond to in-store pricing, a calibration effort could provide a way to translate average online pricing to be more representative of in-store pricing.

There are also several opportunities to expand the types of statistical analyses that are enabled by web crawlers. One could conduct hedonic price modeling methods and include data collected over time. For example, weekly or monthly price and performance data could allow an analyst to include time itself as an independent variable in a regression analysis to allow for statistical models to forecast how price and performance are expected to change in the future. Another approach for factoring time into an analysis would be to use "date added" in the regression analysis. In other words, prices and product features could be time-stamped to the date on which the product was first detected by the web crawler to distinguish between products that have been on the market for several years from those that were only recently introduced, and to determine whether time on the market has a statistically significant impact on price.

Accurate projections of market trends are important to the development and evaluation of effective energy efficiency programs and energy codes and standards. When information is lacking, and future trends in product development are not accounted for, suboptimal programs can leave significant energy and consumer cost savings on the table. The ideas discussed in this paper offer a methodology for taking advantage of web crawling technology to cost-effectively build increasingly powerful datasets that can be analyzed to enhance the effectiveness of future energy efficiency programs.

## References


Desroches, Louis-Benoit, Karina Garbesi, Colleen Kantner, Robert Van Buskirk, Hung-Chia Yang. 2012. "Incorporating Experience Curves in Appliance Standards Analysis." Lawrence Berkeley National Laboratory. September 4.

Flamm, G., Howlett, O., Taylor, G. D. 2012. *Voluntary California Quality Light-Emitting Diode (LED) Lamp Specification*. California Energy Commission, High Performance Buildings and Standards Development Office. Publication Number: CEC-400-2012-016-SD

Young, D., McGaraghan, M., Dewart, N., Borocz, P., Kaser, F., Eilert, P., and Hopper, D. 2014. Leveraging Big Data to Develop Next Generation Demand Side Management Programs and Energy Regulations. ACEEE Summer Study on Energy Efficiency in Buildings 11-332. http://energy-solution.com/wp-content/uploads/2015/01/Leveraging-Big-Data-to-Develop-Next-Generation-Demand-Side-Management-Programs-and-Energy-Regulations_EnergySolutions_ACEEE-2014.pdf.