# CARMA v3.0: Introduction to features, uses, and caveats

Kevin Ummel
CARMA Project Manager
Center for Global Development
July 11, 2012
carma@cgdev.org

**Summary:** This document contains information regarding the appropriate interpretation and use of CARMA v3.0 data. It includes a discussion of features, uses, and caveats concerning geographic data, corporate data, and future data, as well as brief reviews of the CARMA methodology and accuracy of estimated data. Full technical details will be provided in a forthcoming CGD working paper. The blog at www.carma.org will publish this information, along with additions and revisions, over time. If you have further questions, consult the About CARMA page (http://carma.org/blog/about/) and information therein before submitting an email.

## Methodology

Perhaps the most common question from CARMA users is: “Where do the figures on the site come from?” There is a brief answer to this question in the site's FAQ section (http://carma.org/blog/about/faq/). There is also a forthcoming technical paper on the subject. This blog post aims to provide something in-between: sufficient detail for the average user, but not too much.

In an ideal world, the electricity generation, CO2 emissions, location, and ownership of the world's power plants would be regularly published by the appropriate national authorities. Of course, this is not the case. In fact, for the vast majority of countries it is difficult (if not impossible) to find any comprehensive, public information about state of power generators, never mind their environmental impact.[1]

At present, only about 16% of the world's CO2-emitting power plants  regularly disclose CO2 emissions through public databases. These plants are limited to the United States, European Union, Canada, India, and South Africa. Collectively, these databases disclose the specific source of about 36% of global power sector CO2 emissions. A database maintained by the International Atomic Energy Agency also discloses the electricity production of nuclear power plants worldwide. Some databases, like in the U.S., India, and South Africa, report both electricity generation and emissions. Others report only emissions. Some have corporate data, some do not. Some report the location of plants, some do not. Some are exhaustive (covering all facilities in their jurisdiction), some are not. Outside of these sources, information about plant-specific

[1] There are, though, some impressive crowdsourcing efforts to consolidate the fragments of information that do make there way into the public sphere (for example: http://en.wikipedia.org/wiki/List_of_power_stations_in_Malaysia).

performance is fragmented, privately-held, or non-existent. In short, it's kind of a mess.[2]

CARMA's basic task is to consolidate the public data that *is* made available and come up with reasonable estimates for the rest. A private, commercial database maintained by Platts, Inc. provides valuable information about the location, engineering, fuel type, and ownership of effectively all of the world's generating units (though it reveals nothing about actual generation or emissions). This database provides a basis for knowing which plants are reported publicly and which are undisclosed and in need of estimates. It also provides variables that be used to predict the performance of a given plant.

Electricity generation and emissions for undisclosed plants are estimated using statistical models. The U.S. Department of Energy and Environmental Protection Agency publish detailed information about almost all power plants in the U.S. It is possible to process this data to determine what is happening at individual units in particular months. From this data, CARMA constructs a large, detailed dataset of unit-level, monthly performance at U.S. facilities (electricity generation, CO2 emissions, fuel type and consumption, etc.). This dataset is used to fit statistical models that predict how much electricity or CO2 a plant is likely to produce given its size, age, the various technologies and fuels in use, the nature of the electricity grid, etc. The resulting models are then applied to the global database provided by Platts, Inc. to derive estimated performance for power plants that lack publicly disclosed data.

Obviously, there are limitations to this approach. For example, it assumes that the experience and performance of U.S. power plants is similar to those in any other country (controlling, of course, for the various fuel and engineering characteristics that can be observed). The biggest challenge, though, is that utilization rates for plants across time are highly variable. This makes it difficult to accurately estimate the emissions of a given plant in a given year. I cover this issue in more detail in my post on the accuracy of CARMA's model estimates.

While CARMA (or any other effort) will always have difficulty precisely predicting the performance of a given plant in a given year, it *does* do a a few things well:

First, and most obviously, it consolidates the high-quality information that *is* available. This is not a trivial task given that each national disclosure database has its own particular format, standards, and (annoying) idiosyncrasies. And the national databases alone do not provide all the information we desire, which means they must be painstakingly matched against the Platts, Inc. database (and others) to extract the full suite of required information.

Second, even when disclosed data is unavailable, CARMA's statistical models do a decent job of estimating the amount of CO2 a given plant emits for each Mwh of electricity produced (called “Intensity” on the site and given units of kgCO2/MWh). In some ways, the *carbon intensity* is the most important metric, since it allows us to identify those power plants that are the greatest *relative* threat in terms of climate change.

[2] As with most global data, availability tends to be better in developed countries – but not always. It's interesting to note that India and South Africa (both heavily coal-dependent) disclose plant-specific emissions on the web, while Australia (equally coal-dependent) collects the data but will not release it to the public. The same is true of China.

Third, even if CARMA's models cannot precisely estimate total electricity generation or emissions for a given plant and year, the model output *is* likely to be indicative of the long-term performance of a plant. In other words, CARMA's models still do a reasonable job of identifying a plants typical or average emissions over a longer period, even if the performance for any given year is likely over- or under-estimated.

Ultimately, CARMA is a mix of the ideal (disclosed data) and the imperfect (estimated data). The hope is that, over time, better disclosure efforts will tip the balance in favor of the former. Users interested in the U.S. will be happy to know that CARMA's U.S. power plant data come from the DoE and EPA and can be considered high-quality. For facilities outside the U.S., it is possible to check the disclosure status of a given plant by downloading the associated .csv file from the site and finding the "dis" variable in the output. This variable indicates one of the following situations:

dis=0: No data disclosed
dis=1: Electricity generation disclosed
dis=2: CO2 emissions disclosed
dis=3: Electricity generation and CO2 emissions disclosed

**[Note: This feature is in the process of being added to the website and .csv downloads]**

---

**Estimated data accuracy**

Although CARMA incorporates all known major public disclosure databases, the majority of the site's data is necessarily estimated using statistical models. This will hopefully change in the future as governments and companies become more open about the source of global warming pollution, but for now estimates are unavoidable. So, how accurate are CARMA's model estimates?

This question is addressed in detail in a forthcoming technical paper describing the CARMA methodology and results. Here I want to share some of the main findings and highlight important caveats related to use of the data.

As detailed elsewhere, CARMA's models are fit to a high-resolution dataset of U.S. plant performance. The models then predict the electricity production and CO2 emissions of plants outside the U.S. for which publicly disclosed data is not available. As part of the CARMA technical paper, an analysis was undertaken to estimate the likely accuracy of the model output.

Overall, the models do a better job of predicting the carbon intensity of a given plant (kg CO2 per MWh) and have more difficulty accuracy predicting total electricity generation. For example, it is estimated that, for emitting plants with estimated values, CARMA reports CO2 intensity that is within 20% of the true value about 60% of the time. But for electricity generation, the reported value is within 20% of the true value only about 40% of the time.

Why is predicting the amount of electricity generated by a given plant in a given year so difficult? The short answer is that utilization of many plants jumps around from year to year (i.e. high inter-annual variability) for reasons that cannot be easily observed or modeled. For example, the CARMA technical paper analyzes how annual generation changed between 2009 and 2010 for ~5,000 U.S. power plants that showed no change in engineering characteristics. Nearly 50% of the plants saw annual generation change at least 20% between 2009 and 2010 and about 30% saw a change of at least 40%. Remember, the variables that CARMA's models are able to use have not changed for these plants – but generation is still jumping around from year-to-year. This variability makes it fundamentally difficult to detect clear patterns or "rules" that the models can use to precisely predict performance when public data is not available.

When we consider these difficulties, the CARMA models are actually performing reasonably well. For example, an "ideal" model, given the range of variables available to CARMA and accounting for inter-annual variability, would likely predict annual generation to within 20% of the true value in about 55% of cases. The evidence suggests that the CARMA v3.0 models currently achieve that level of accuracy for slightly more than 40% of plants. And whereas an ideal model could be expected to be within 40% of the true value for about 70% of plants, CARMA does so in more than 60% of cases. Overall, that's pretty decent model performance.

It's also clear that accuracy depends on the type of power plant in question. In general, larger plants are easier to predict than smaller one. And coal power plants – owing to their predominant and more consistent use as base-load providers – should enjoy greater model accuracy than other fuel types. Conversely, smaller and/or gas- or oil-based units are likely to see higher prediction errors. Hydroelectric power plants are a mixed bag since performance in any given year is highly dependent on local weather conditions that are not observed by CARMA's models.

On the plus side, CARMA's estimates *can* be fairly interpreted as reasonable long-term performance metrics. CARMA's models show no evidence of systematic bias, so while estimates for any *particular* year may exhibit significant error, the long-term performance of most plants is likely consistent with the model predictions. This is especially true of larger plants. Measures of typical, long-term performance for larger facilities (existing and planned) are, perhaps, the most relevant information for many real-world applications of CARMA. In addition, prediction of CO2 intensity – an equally useful metric for many CARMA users - is shown to be quite feasible and exhibits relatively low error.

---

**Geographic data**

CARMA v3.0 greatly improves both and scope and quality of geographic information provided for individual power plants.

Basic information like country, state/province, and city comes from a proprietary, commercial database of global power plants. Similar data is provided for U.S. facilities by the Department of

Energy (DoE). This raw data is processed with an algorithm that cleans and standardizes the data and conducts a "fuzzy string" match against the open-source GeoNames place names database. The algorithm attempts to extract maximum geographic information; in some cases, it is able to add information not found in the raw data. I believe this makes CARMA v3.0 probably the most extensive public geocoding of global power plants to date.

CARMA's geocoding algorithm attempts to return the continent, country, state/province, county/district, city, and postal code for each plant. Data coverage is universal for continent and country and nearly so (>95%) for state/province. A further 80% of plants have been assigned a city, 40% a county/district (i.e. secondary region), and ~16% a unique postal code.[3]

CARMA users are often interested in pin-pointing the location of facilities, usually for the purposes of modeling pollutant dispersal or making high-quality maps. This requires specific geographic coordinates. Coordinate data from the DoE and EPA is used to provide high-resolution coordinates for all plants in the U.S. Outside the U.S., the same datasets that disclose emissions or power generation sometimes report coordinates, too. In addition, many large facilities have been manually geocoded using public sources (usually Wikipedia). All told, 12% of facilities responsible for about 40% of current electricity and emissions are assigned high-resolution coordinates.

When high-resolution coordinates are not available, CARMA v3.0 provides the coordinates for the associated city center, as given by GeoNames. An additional 70% of plants are assigned these approximate coordinates. Comparison of approximate and precise coordinates for plants with both suggest that the approximate coordinates have an average spatial error of about 7 km. When downloading a .csv file from CARMA.org, a variable called "crd" is included to indicate if the given coordinates are approximate (crd=1) or precise (crd=2).

**[Note: This feature is in the process of being added to the website and .csv downloads]**

The CARMA website reports aggregate totals for the geographic entities previously mentioned, as well as counties, congressional districts, and metro areas for the U.S. The definition of a "metro area" has changed in v3.0 and now reflect the borders of "combined statistical areas", as determined by the OMB (http://en.wikipedia.org/wiki/Combined_statistical_area). For users of CARMA's API, it is important to note that all regions in CARMA v3.0 (excluding congressional districts and metro areas) now have unique, permanent identifiers that match those used by GeoNames. For example, the Australian state of New South Wales (http://carma.org/region/detail/2155400) has region_id=2155400 (specified in the URL), which matches that used by the GeoNames API. This allows the two databases to be easily linked, if desired.

The regional totals provided in CARMA v3.0 are simply the aggregate electricity production and emissions of all geocoded facilities within the borders of the region in question. The one exception is cities. The city totals (for example, Madrid: http://carma.org/region/detail/3117735) are the aggregate of all plants with precise or approximate coordinates within 100 km (~60

[3] Secondary region (i.e. district/county) information is not currently provided to the public but is available upon request.

miles) of the city in question. CARMA v3.0 provides such totals for capital cities and those with population greater than 50,000 – more than 13,000 cities worldwide.

It's also worth noting that CARMA's algorithms attempt to ensure accurate country totals for electricity generation and, for most countries, CO2 emissions. National totals from the DoE and International Energy Agency are used. There may be discrepancies in some cases. If you notice any, please let me know.

---

**Corporate data**

As in previous releases, CARMA v3.0 includes information about the electricity production and emissions of corporate entities involved in power generation. Every plant in the CARMA database is assigned to a company. The vast majority of that information comes from a proprietary, commercial database underpinning CARMA. In some cases, corporate ownership of U.S. plants is also provided by data from the Department of Energy.

Power plant ownership is quite complicated, and there are often multiple layers of ownership between the immediate plant operator and the ultimate owner. CARMA attempts to report the *ultimate* owner whenever possible (i.e. the highest entity in the corporate hierarchy), relying largely on information from a private data supplier. When the parent company cannot be identified for a plant, the operating company (often a utility) is reported instead.

It must be stressed that maintaining accurate corporate information is extremely challenging. CARMA's data suppliers are the best in the business, but even they cannot guarantee accuracy, especially outside North America where corporate hierarchies are less evident. There is also the problem that many large facilities are owned by multiple entities and it is not possible to track ownership shares worldwide. For all these reasons, the company data in CARMA should be considered a reasonable "best guess" of the primary entity ultimately responsible for ownership or operation of the plant.

For example, the generating units at the Scherer Plant in Juliette, Georgia (the largest CO2 emitter in the U.S.) are jointly owned (to varying degrees) by seven different corporate entities (http://en.wikipedia.org/wiki/Plant_Scherer#Operator_and_ownership). Based on this information, one could reasonably consider Oglethorpe Power Corporation (http://carma.org/company/detail/14613) the primary owner. On the other hand, the plant is actually operated on a day-to-day basis by Georgia Power (also part owner), which is, in turn, a subsidiary of Southern Company (http://carma.org/company/detail/18979) – as is Gulf Power, also part owner of the plant. CARMA's data suppliers, and hence the CARMA database, report the parent company as Southern Company (http://carma.org/plant/detail/40093). I use this example to illustrate the potential complexity of ownership arrangements and dispense the necessary "grain of salt". To be fair, this is an especially complex case.

Although electricity and emissions data are available for multiple points in time (e.g. 2004, 2009, and the "Future"), the company assigned to each plant is based upon the most recent information available. For example, the CARMA v3.0 release in July, 2012 uses corporate ownership data thought to be current as of March, 2012. Notice that this makes company totals for earlier points in time (e.g. 2004 and 2009) possibly subject to error, since ownership could have been different at that time. In other words, the 2004 company totals for Southern Company, for example, are the aggregate 2004 electricity production and emissions of plants that Southern Company is *currently* listed as owning (not the plants they *actually* owned as of 2004). In general, the 2009 company totals are more likely to be accurate than 2004, especially in developing country power markets where changes are occurring rapidly.

CARMA also assigns a “Parent Country” to each company in the database. At present, this is based solely on an analysis of where the plurality of the company's associated generating capacity is located. In the vast majority of cases, this yields an accurate result. But for multinational corporations, this can result in errors. For example, RWE AG (http://carma.org/company/detail/17285) is a German company, but the United Kingdom contains a plurality of its generating capacity. Looking up the headquarter country of every company is not feasible, so if you come across errors of this nature, please contact me so I can rectify for future releases.

Finally, it is worth reiterating that CARMA's company totals simply reflect emissions from associated power plants. Occasionally, one will find companies that have both power plants and other business activities. For example, major oil and gas companies like BP (http://carma.org/company/detail/2314) often own generators, and their operation is reflected as best as possible in CARMA's company totals. But these figures do *not* reflect the full extent of CO2 or other greenhouse gas emissions from BP's wider activities (drilling, refineries, pipelines, etc.).

---

**Future data**

One of CARMA's rather unique features is the inclusion of data about the future. The v3.0 data contains entries for year 2004, year 2009, and the “Future”. All three points in time are displayed on any of the detail pages at CARMA.org (for example: http://carma.org/plant/detail/44204).

Many CARMA users are interested in information about future developments in their area. Where are plants being constructed or planned? Where are existing plants being expanded? Which companies are likely to see their emissions rise the most? A proprietary, commercial database underpinning CARMA provides information that can help answer these questions.

First, a word of caution: The underlying database that provides information about future developments is only as good as the state of public information around the world; it reports what companies and plant builders have actually divulged. In some cases, the reported plans may be concrete and comprehensive. In other cases, they may be tentative and incomplete. There is no way of knowing which is which. In short, the “Future” figures in CARMA must be interpreted with caution. I want to briefly show some examples of how to (and how *not* to) use this information.

Let's consider individual plants. The simplest "Future" case is the construction of a new power plant. Such plants are included in CARMA with the signifier "(Planned)" appended to the plant name: http://carma.org/plant/detail/74556. We can see that this plant was not in operation in 2004 or 2009, but data are included for the "Future" period. "Future" refers to any point in time after 2009. So, a planned facility might have entered commercial operation last year – or it may not go into commercial operation for a decade or more. Information about start dates is sometimes included in CARMA's input data, but (unfortunately) cannot be released to the public due to proprietary data restrictions.

In the case of a planned plant, the "Future" data are simply a model estimate of plant performance once commercial operation begins, based on engineering specifications. The best way to search for planned plants via the website is to use the Dig Deeper tool (http://carma.org/dig/) and sort a given locale's power plants using the "Future" radio button on the right side.

In some cases, future plans include capacity expansions at an existing facility. The Taichung plant in Taiwan is a case in point: http://carma.org/plant/detail/44204. We can see that electricity generation and CO2 emissions jump up in the "Future" period compared to 2009. If a plant shows no change in data between 2009 and the Future, that is indicative of no planned capacity expansions (or retirements). If the two sets of data are different, then some planned change(s) is expected, though the date(s) of the change is uncertain. The "Future" data are, again, a model estimate of how the plant might operate after the alterations are completed.

"Future" data are also available for geographic regions and countries, though they should be treated carefully. In both cases, the figures simply report aggregated "Future" electricity production and emissions from all associated power plants. We do *not* know if the "Future" totals reflect plants planned for operation in the next 5 years or 20 years. Nor do we know if the reported future plans are exhaustive or a small sample of what will actually occur. Looking at the totals for Hebei Province in China, for example (http://carma.org/region/detail/1808773), we see that "Future" CO2 emissions are significantly higher than 2009 emissions. But whether this increase occurs by 2015 or 2025 – and whether actual emissions go even higher – is impossible to tell from CARMA's data alone.

Overall, CARMA's "Future" data is most helpful in revealing planned power plants that were not in operation as of 2009 ("new builds") and a reasonable estimate of their likely electricity production and CO2 emissions. The "Future" data can also be used to identify the likely effects of capacity expansion (or retirement) at existing facilities by comparing the 2009 and "Future" data. The "Future" data are far less helpful when looking at region and company totals. In those cases, uncertainty about the timing and comprehensiveness of future plans make the totals difficult to reliably interpret. Users should be aware of these limitations and recognize that "Future" data reported by CARMA are by no means exhaustive projections or predictions.