Conserving Energy for a Better Tomorrow

An eSC Initiative

Authors:

Luis Alfredo Riviere

Ishaan Lodhi

Marko Masnikosa

Executive Summary

This project for eSC, an energy provider in South Carolina, aimed to forecast energy consumption for July 2025 and identify strategies to reduce consumer energy usage amid rising temperatures. Using county-level data on housing, weather, and energy usage, we developed predictive models after extensive data preprocessing and analysis.

Gradient boosting outperformed other models, achieving an R-squared of 0.76 and RMSE of 0.39, effectively capturing the complex interactions between static and dynamic features. Separating data by climate zones enhanced accuracy and provided tailored insights. Simulations revealed nonlinear energy surges in hot, humid climates, emphasizing the need for targeted energy strategies. Actionable recommendations included improving HVAC efficiency, upgrading appliances, and optimizing solar panel systems to manage future energy demands.

Introduction (scope/context/background)

Business Questions addressed

We represent eSC an energy provider in the state of South Carolina. Due to rising energy usage and global temperatures, we want to be better equipped to handle increased power demands. To do so we will look at county-level energy usage data for households and try to extrapolate results for the upcoming summer month of July 2025 so we can meet power production targets. We were also interested in ways we could reduce energy consumption in general on the consumer size.

Data Acquisition, Cleansing, Transformation, Munging

Our data consisted of the following sets:

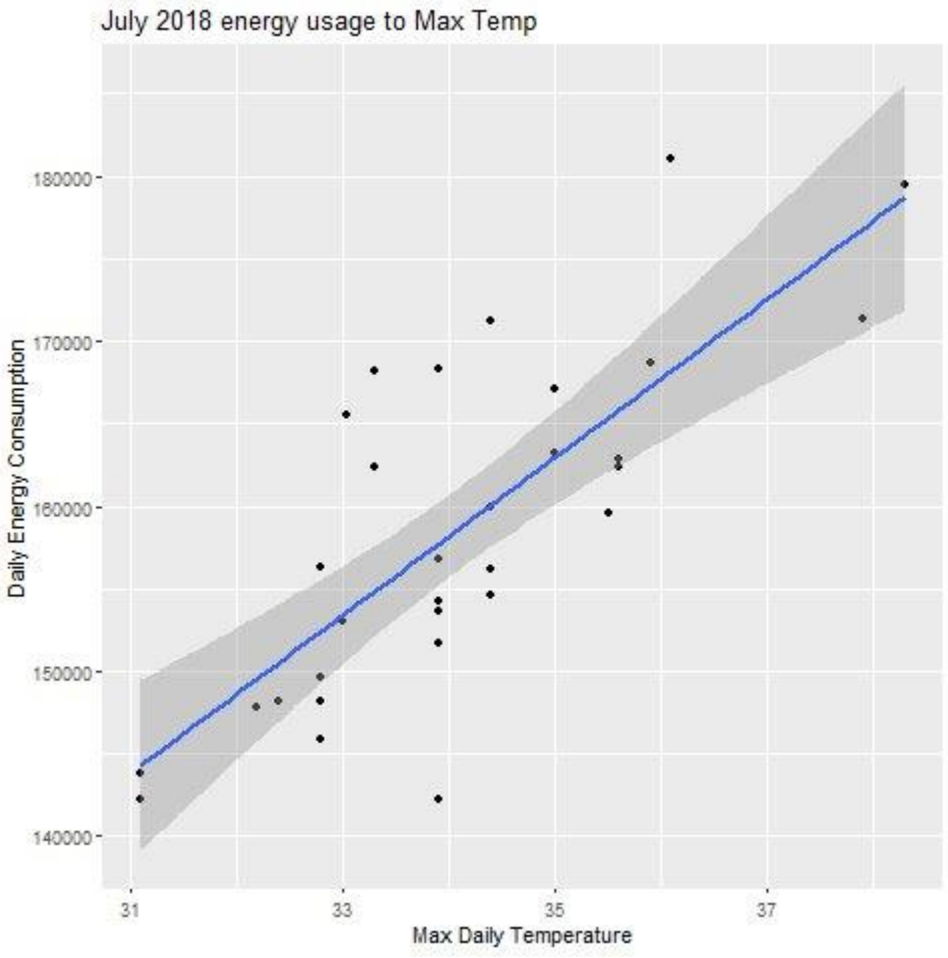
Housing data: Consists of over 5000 households and more than 170 static features including but not limited to, location data such as latitude, longitude, and climate zone, insulation type, domestic appliance type, square feet, garage size, ac usage and many others.

Weather data: Hourly weather data by county divided into 6 distinct features. Dry Bulb Temperature (absolute temperature in an environment with 0% humidity or additional pressure), Relative Humidity, Wind Speed, Global Horizontal radiation (total amount of solar radiation that falls on a horizontal surface on Earth), Direct Normal Radiation (the amount of solar radiation that reaches a surface that is perpendicular to the sun's rays, per unit area), Diffuse Horizontal Radiation (the terrestrial irradiance received by a horizontal surface which has been scattered or diffused by the atmosphere.)

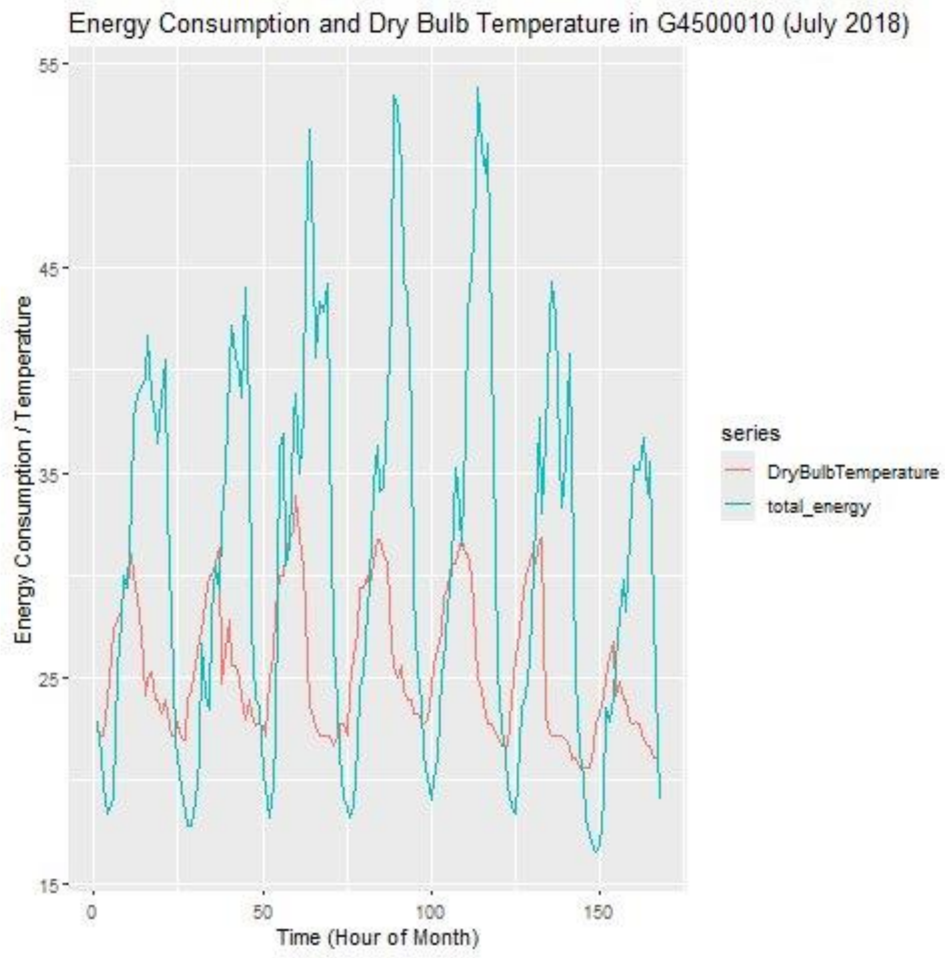
Energy Data: Hourly energy data for each building, includes electrical and natural gas usage for different appliances for each building in the data set.

The data available for this project spans the whole of 2018 and early 2019. Altogether, the data was over 50 million rows, with over 200 features. Marko started gathering the data and blending it together from the various sources, reducing the dimensionality of the overall dataset by dropping the features that only consisted of a single value as this would not help to explain the variance in the dependent variables. Using low variance filtering, we were able to drop 70 features, mostly from the static features. The remaining dataset was still too large to fit into memory on most machines, so the dataset was broken into 10 parquet files that were stored for later analysis. The energy consumption figures provided were specific to the energy type (electricity, natural gas) and application (HVAC, fridge, grill...), so we had to calculate the total energy consumption per building. All provided energy figures were in kWh, so calculating total consumption was straightforward. PV energy generation was provided as a static feature, not hourly, so it was not included in the calculation. Once the data was scraped, we started merging. First, we merged the static data with the consumption data on the building ID and then we merged that data with the weather data on time. We dropped unnecessary columns that came with the energy database. Once formatted, we worked on imputation. The first step was finding out the variance in the different static features. If features had no variance, we dropped them since they would not assist in model creation. We then moved on to filling out the missing values. When sensible we used back and forward fill function to look at past and present values by building id to fill. This is assuming that status features like size, insulation type etc. Will remain consistent in the same building. If there were other N/As that could not be explained by a None, we then proceeded to drop them, 8% of rows. Our next step was factoring. To extract useful information for model training we had to turn all string and char outputs into numerical and to do so we had to factor first. This involved researching all relevant columns and ranking them by electricity consumption efficiency. To further explain this here are some examples. When looking at insulation we factored them with 0 being the worst rated insulation and then moved on to more efficient insulations, the same was done with all other string columns, we researched energy efficiency and ranked them as so.

Descriptive statistics & Visualizations



Above is a graph depicting the relationship between daily max temperature and daily energy consumption. As we can see there is a significant correlation between energy use and temperature. Although this is to be expected it was important to verify this relationship before moving on to model development.



Here we have another plot displaying this same trend but on an hourly basis. As we can see, the temperature reaches its peak a few hours earlier than the peak energy consumption for that day is reached. We hypothesize this is due to the time it takes for building temperatures to rise due to the insulation of the buildings.

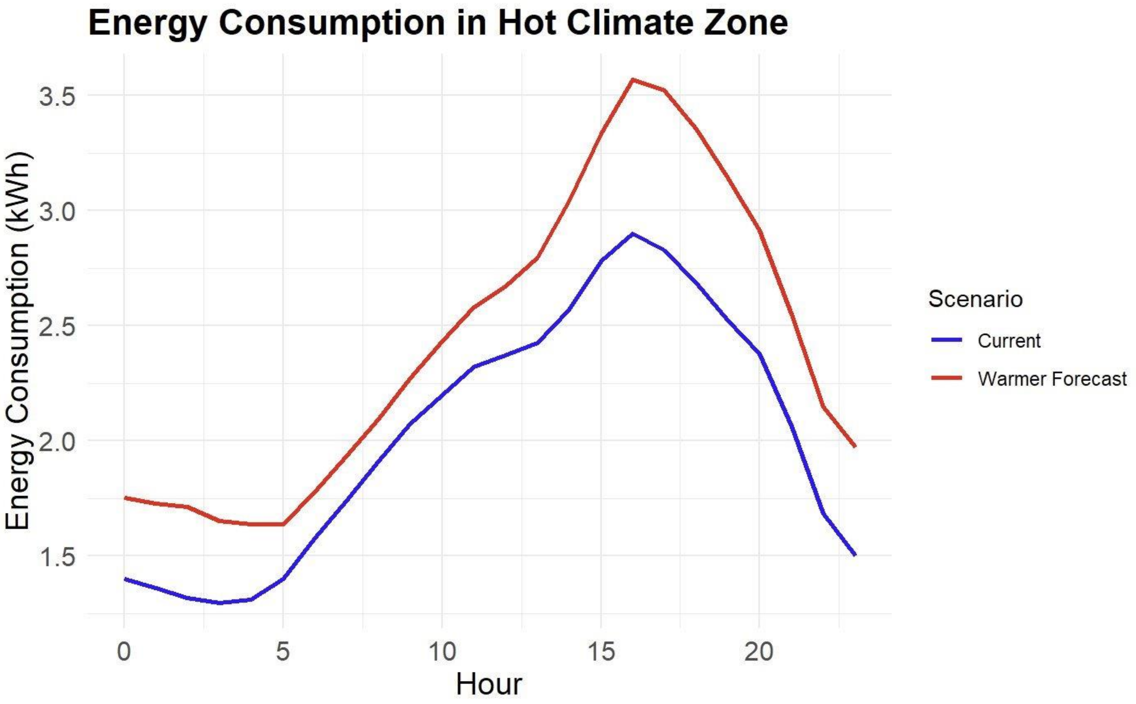
Use of modeling techniques & Visualizations

Linear: Our first model was made for simply exploratory purposes. We developed a linear model that took the total energy consumed per building hourly as the dependent variable and the weather and static housing data as the independent variables for July. The result was expected as the data is not linear. We ended up with a model that had an R-squared value of 0.46. Due to this low explanatory power, we dropped it right there and then.

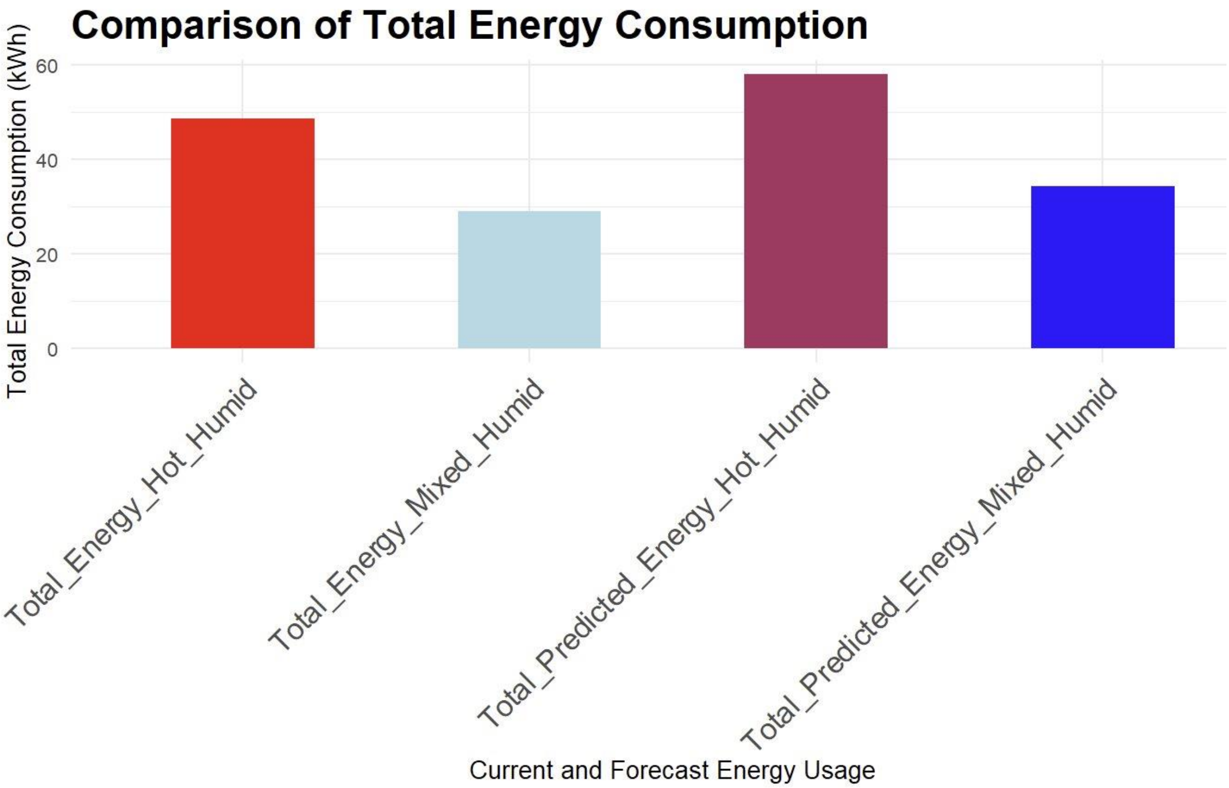
ARIMAX: Although a vast improvement in precision, it did not consider static features. For the ARIMAX modeling approach, we used the forecast and tseries packages. For this model, only July 2018 data was considered as this would be the closest example of a hot summer month that was available to model on. This data was then grouped by county and time to improve model runtime and to better understand the differences between geographic locations. The Augmented Dickey-Fuller test gave a strong indication that the data at this level, for each county, was stationary, which made modeling easier. While the ARIMAX did capture the day/night cycle of the underlying energy consumption and weather features, the error was still great as the static features were not able to be incorporated due to them not having a temporal aspect. From our exploration, we saw a pattern of energy consumption spiking hours after the temperature hit its peak for the day and that difference in timing and duration was more adequately captured when the static features were incorporated in other models. We also briefly tried TBATS which would apply an exponential smoothing to the time series, but according to the author of the package himself, the TBATS model does not consider exogenous variables, so that model was not pursued further.

Panel Regression: our search for a model that could account for the data's nonlinear nature and could account for static features took us to panel regression. Panel regression models are used to analyze data that has both a time and a cross-sectional dimension. We used Principal Component Analysis to narrow down the number of relevant features. We started dropping the features that explained the least amount of variance. We once again used the same data from our linear model, this time with fewer features. Despite what we assumed where improvements the generated R-square was even lower than our linear model. Which also meant we dropped it on the spot.

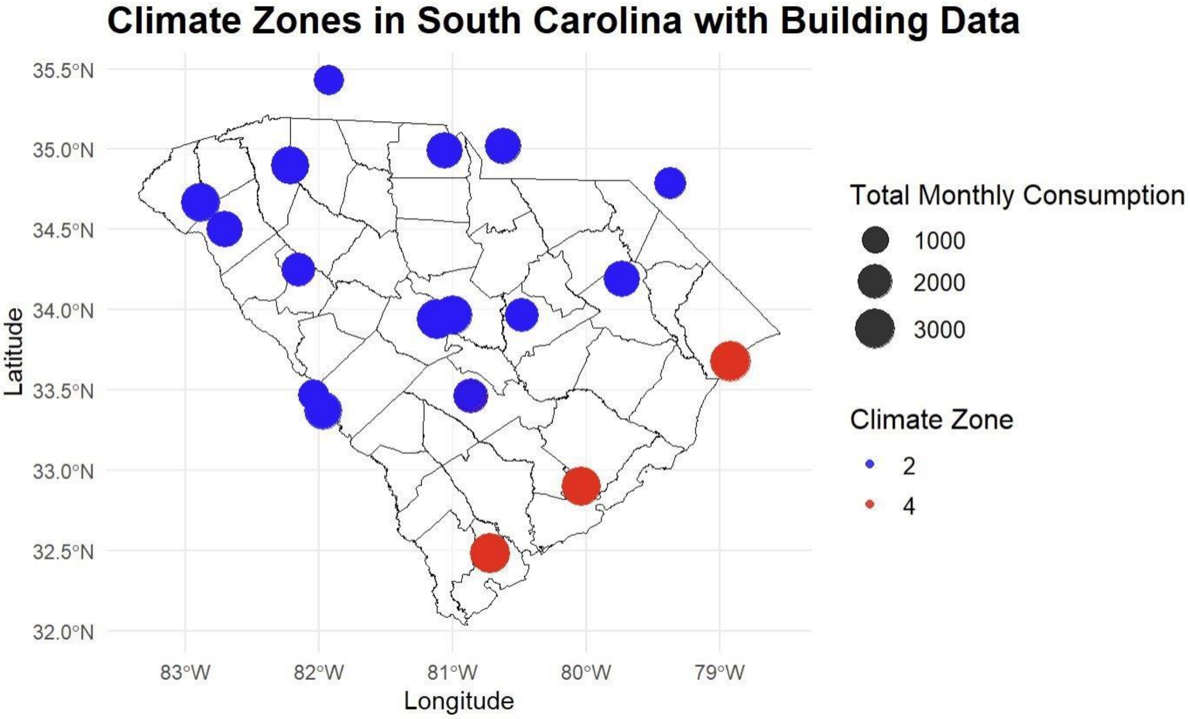
Gradient boost: Given the complexity of the interaction between static features and energy consumption, we shifted to ensemble learning methods and selected gradient boosting as the best approach. Gradient boosting models employ decision trees as base learners; however, they differ from other ensemble approaches, such as random forests, in their iterative learning process. Specifically, gradient boosting builds decision trees sequentially, with each tree trained to minimize the residual errors of previous trees using a gradient descent optimization approach. This enables the model to incrementally correct its errors, efficiently reducing bias and variance. While the individual trees in the ensemble are weak learners, their combined effect produces a robust predictive model with high explanatory power, well-suited to capturing subtle interactions between static variables and energy consumption. This method led to a significant increase in model performance. The gradient boosting model achieved a much higher R-squared compared to any previously tested models. To further enhance model accuracy, we divided the data by climate zone—a highly effective step in capturing regional fluctuations in energy usage. This adjustment substantially improved the model's explanatory power, resulting in a final R-squared of 0.76. For the predictions, we used an 80/20 train-test split, yielding an RMSE of 0.39 on the test set. Finally, we simulated a scenario by increasing the Dry Bulb Temperature in the weather parameters by 5 degrees and reran the model to forecast energy consumption for a hypothetical July with elevated temperatures. These predictions provided valuable insights into the potential impacts of increased temperatures on energy demand.



Above, we see a graph showing the current versus forecast energy use with warmer temperatures for July. This graph is only for the Hot Humid climate zone which is one of the two included in the dataset. To note is the fact that the biggest difference in the two graphs is during peak. This suggests that the increased energy depending on temperature is nonlinear as it increases rapidly at higher temperatures. Which further emphasizes the importance of the work we are doing since energy use will not only rise with higher temperatures but after a certain point rise exponentially.



This bar plot compares current and forecast energy usage in the two climate zones in the data, Hot Humid and Mixed Humid for July, both present and forecasted with an additional 5 degrees. As we can see there is a stark difference between the two. Even with an additional 5 degrees the Mixed Humid climate zone consumes less energy than what the Hot Humid does now. This corroborates our assumption that energy use rises exponentially as temperature increases and not at a steady rate.



In this final map we see where the buildings in our data set are located, as well as their climate zones. We see here how most of the buildings in the Hot Humid climate zone are located on the coast of the State of South Carolina while the Mixed Humid are located further inland. We know that the Hot Humid Climate Zone accounts for most of the energy consumption in the month of July. Although higher temperatures in this climate zone might explain the significant increase in energy consumption it is also worth noting that this area of the state could be seeing additional traffic in the form of tourism during this month that could be skewing the result. Either way, eSC has to be prepared to meet this energy demand.

Actionable Insights / Overall interpretation of results

When looking at actionable insights we looked at the PCA features with the highest impact, out of these we further investigated the ones that the consumer could have some level of control of. For example, we did not select insulation since changing the insulation of a house is a huge feat, but we did look at other features such as appliance type. We calculated the mean value for each factor in a feature and then looked at where significant energy spikes or drops. This left us with three features we thought were worth sharing. The first was heat pumps. According to our data the presence of a heat pump is the singular change that most affects HVAC efficiency. The second change of most impact was refrigerator type. We found that refrigerators that had an efficiency rating of less than 6.7 consumed a significant amount of more energy. Slight improvements in the newer models were not as great. An EF of 6.7 was the cutoff. Our final and we could argue most interesting observation was solar panel system size. In this case we are referring to size in terms of energy that the system can handle, not size of the solar arrays themselves. We found that solar panel systems had a capacity greater than 7 kWh were actually correlated to higher energy expenditure. This question requires further investigation. So far, we have a few hypotheses to explain this. The first is that households that have arrays of this size might simply have a larger footprint and consume more energy in general. The second is that this is simply the most efficient size for a solar panel system in the state of South Carolina given its amount of sunny says and day length and the third is that there might be a psychological effect in which if a household has a larger solar panel system they are more likely to partake in energy consumptive behavior.

Moving Forward

Regarding the future of the model the next steps are, deploy the model on live data. That way the model can constantly generate predictions and learn from error as time catches up to its predictions. Leading to increased accuracy. Our other recommendation to eSC to focus on marketing initiatives, financial incentives, or work with the local government to enact policy that promotes the adoption of the previous measure we found more impactful.

Conclusion

This study provided eSC with a robust framework for predicting energy consumption under rising temperatures and identifying consumer-side interventions to reduce energy demand. The gradient boosting model demonstrated its capacity to accurately capture the complex interplay of static and dynamic features. By identifying key factors such as HVAC systems, appliance efficiency, and climate-specific behaviors, this analysis offers actionable strategies for optimizing energy usage. Future steps include deploying the model with live data to enhance predictive accuracy over time and leveraging the findings to inform policy, financial incentives, and consumer education. These efforts will enable eSC to meet energy production targets and support sustainable energy practices in South Carolina.