Data Science and Ecological Restoration: 4 Steps to Action with a Real-Life Case Study

Introduction

I used to be on a track quite different from data science. I went to a very good school for Forestry and Conservation. And even in one of the best schools in the world for the topic, using computers was a touchy subject. In the handful of courses that were oriented around teaching us excel, geographic information systems, even word processing, there was always a feeling that loomed overhead, with many students giving each other glances across the computer lab: *I picked this major so I could be outside! Not stuck behind a screen!*

Even once I’d graduated, the feeling persisted for some people. My boss would frequently echo frustration about not being able to follow along with us in our field work, and colleagues would frequently struggle to use computers beyond basic office software.

It was never quite that way for me however. I loved computers, and still do, which is the main reason why I started dabbling with exploring data in Python. This blog shows the results of my first foray into data science with actual, real-world data, and the insights that it brought to the project, and hopefully can provide any other environmentalists who struggle with technology a new insight for how this tool can be used.

And if any seasoned data scientists are looking at this, I hope that instead it serves as a little introduction into ecological restoration, and drives at least one more person towards a greener future.

Ecological Restoration: What is it? How’s data science supposed to help?

The goal of any ecological restoration project is to start with heavily modified land – wasteland, old quarries or fields – and to establish a living system within it similar to what existed before human intervention. This benefits both natural populations – Habitat loss is one of the [leading drivers](https://www.nature.com/articles/536143a) of species extinction – and human populations; aside from it being generally pleasant to have forests, they also serve to clean the air and capture useful water, and many other [ecosystem services](https://en.wikipedia.org/wiki/Ecosystem_service).

The data I’m going to examine is from an ecological restoration project in Colombia, run by the environmental NGO Fundación Natura (seriously, [check them out](https://natura.org.co/)). The project is aiming to increase the coverage of [Andean Cloud Forest](https://www.natureandculture.org/ecosystems/andean-cloudforests/) by tree planting in unused agricultural land. The particular plot I will be examining features roughly 2700 trees planted in late 2019. The plantation has seen moderate success but suffers from high mortality and a limited budget. The data collected measures plant biomass and health indicators, taken roughly 3 times a year, with interruptions in the process courtesy of your local global pandemic.

First things first: This is not big data. This won’t feature machine learning. It is, however, real on-the-ground stuff that can improve outcomes and save money, but also highlights the difficulties of applying raw numbers to living organisms.

So what can the data tell about your restoration project? Here are some of the key insights that I derived:

**Insight 1: Evaluate the general status of your project by measuring key performance indicators.**

The purpose of a planting operation is primarily to establish itself by surviving and accumulating biomass. You can measure this with plant Height, Basal Stem Diameter, and Crown Coverage. For non-ecologists, you want biomass because it tends to represent the plant’s ability to thrive and outcompete its neighbors. In this case we want the plants to outcompete grasses and any invasive species that are hanging around.

Relative plant health is a companion indicator that shows the proportion of biomass being affected by various plant conditions that we’ll go over later. In this case a lower value is better.

Figure 1: Biomass indicators over time:

You can see that biomass is increasing over time, but plant health is also deteriorating.

Similarly, a second metric for success is plant survival and health. You want plant survival to be high, mortality to be low, and any plant ailments to be low as well.

Figure 2: Survival and health indicators over time.

As you can see, we initially had some really bad survival numbers that have gradually improved. Plant health on the other hand, is inconsistent. You can see that we actually only have 3 health indicators that show any kind of consistency over time: Coloration, Low Vigor, and Herbivory. These 3 values are likely tied to broad ecosystem factors that are difficult to change, but easy to factor into logistics calculations. The other 5 on the other hand probably represent acute events that you may or may not be able to influence by your decisions.

**Insight 2: Drill down on specific pain points.**

In the previous section we identified some variables that are negatively impacting our success metrics. It would take too long to examine each one in this post, so I’ll focus on one in particular: Plant predation by wild guinea pigs.

Image 1: Guinea Pigs. A cute problem.

Domestic guinea pigs originated in the Andes, and wild populations can be found widespread throughout the area. They like to live in tall grass, and they love to eat tender vegetation. In this case, the bark of our newly planted trees is soft and full of simple sugars that these critters find irresistible. If a guinea pig eats the whole way around the stem of a tree (A process called girdling) it can kill the plant.

What can the data tell us about their behavior? We know that a guinea pig has munched on a plant because it leaves behind stripped bark that will form into a scar if the plant survives.

Figure 3: A Guinea Pig’s Favorite Lunch (And breakfast, And dinner)

As you can see, guinea pigs don’t seem to be very picky with their food. However, two species are overrepresented in the data, one by the sheer amount of trees with bite marks, and the other by the proportion of individuals of that species that were eaten out of the amount that we planted. Maybe this is a list of plants to avoid as we replace dead individuals with new ones.

One thing to keep in mind here is that if a plant dies, we no longer track it in our dataset. Secondly, we also don’t keep track of the exact cause of death of a plant. Thinking about it differently, these plants represent the plants that survived the most after guinea pigs pass through. This way we avoid a potential source of survivorship bias, and instead this becomes a list of plants that we know can handle the pressure put on them.

What about the guinea pigs themselves? We can actually track their behavior by mapping where predation incidents are the highest. The plantation is structured into groups of 56 plants called modules, and each module is tagged with geographic coordinates. If we map the incidence of guinea pig attacks, it looks like this:

Figure 4: Geographic Distribution of Guinea Pig Attacks

The guinea pigs seem to like the center of the plot as opposed to the edges. Has this always been the case?

Figure 5: Geographic Distribution of Guinea Pig Attacks Over Time

If I was a real GIS expert, I could have overlaid this on top of a map or aerial photograph of the actual reserve. For now, please bear with me with this abstraction.

Spreading the attacks out over time appears to reveal a new pattern: There appears to be a cyclical nature to when guinea pigs will predate on the trees. What could be causing this? One thing that we haven’t examined so far is the age of plants: In this particular plot, we are contractually obligated to replace any dead trees with new ones. These replanting events were done just before every data gathering event, and if we add that data to our previous figure, it looks like this.

Figure 6: Geographic Distribution of Guinea Pig Attacks Over Time, by Time of Planting

Notice how each period where there was a surge in predation is now spread to each of the individual cohorts that we planted. The simplest conclusion we can draw now is that guinea pigs love young plants, and they tend not to revisit plants they’ve already eaten. It also seems to take them a bit of time to discover when we’ve introduced new plants.

What actions could we take to mitigate some of these effects? Any plant mortality has a cost measured in dollars and man hours, so we want avoid it as much as possible. It might therefore be worth it to invest more in protecting newly planted individuals, either by fencing them off or by chemical means (maybe a nice application of capsaicin will scare them off). If we can succeed in protecting the young plants for long enough, we may reduce the number of plants that we need to replace.

**Insight 3: Find patterns that you can influence by your actions.**

Seeing the geographic distribution of one metric allowed us to come to interesting conclusions and is a dynamic and intuitive way of seeing real-world data. What might we see if we apply it to the other variables?

Figure 7: Geographic Distribution of Biomass Metrics. Darker is Better

Figure 8: Geographic Distribution of Survival and Health Metrics. Lighter is Better

Looking at the charts, there appears to be two patterns that appear fairly regularly: The plants grow largest in the southern half of the plot, while plants seem to grow healthiest around the edges, except for the north edge. You may see other patterns as well, but I’ll be focusing on these two patterns.

You have to wonder, what is it about the makeup or conditions in the southern block and edges that is responsible for this effect?

Figure 9: Geographic Distribution of most successful plants

Figure 10: Geographic Distribution of least successful plants

Our most successful plants are simply the ones with the highest survival rate. Turns out that they survive the most near the south, and less to the north. If you look at the least survival plants, the relationship is less clear, but some of them do indeed seem to have the opposite pattern: more of the least successful plants have higher survival rates up north.

This could be because of any number of environmental variables, but at the very least it opens up a path for action: If we concentrate the plants that do best in their own respective zones, we may be able to maximize plant survival while keeping species diversity. If we can determine which environmental factors determine this pattern, we could even incorporate it into future projects, but this will most likely require more study.

Finally, what about the edge effect? Edge effects are a commonly studied phenomenon in ecology: Plants tend to behave differently when they have to grow in a forested area vs out in the open, and this has lots of ramifications in interspecies interactions, from plants to animals. Around this plot, with the exception of the northern edge, there is a forest. On the northern edge there’s a path and then more planted trees. Proximity to the forest seems to have some beneficial effects.

Figure 11: An aerial photograph of the planted area. Notice the forested edges.

**Insight 4: Use context to avoid conclusions that seem too good to be true**.

You may have noticed in the previous section that the geographic distribution for competition has a very distinctive pattern.

Figure 12: Geographic Distribution of Competition

That’s a pretty sharp grouping of values, right? Too sharp. A visual glance may not suffice but fortunately the original dataset has a “Comments” column that may reveal some insight. Turns out that a frequent comment made is “The plant is swamped by grass because we haven’t cut the grass in this area yet, otherwise the plant is fine”.

While these comments don’t correlate perfectly with the plants marked as having competition, this is enough noise that the data in that column shouldn’t be trusted. At the very least the way competition data is gathered should be re-evaluated.

Conclusions:

In conclusion: You can use data science for a whole range of questions you may have about an environmental restoration project, as long as you’re collecting enough data. The more specific your questions about the data, the more complex the answer may be, and the harder it is to draw conclusive answers.

Regardless, every step of the way raised some interesting patterns and pertinent questions. For example targeting which species we plant and where could be a cheap way to improve outcomes in the future, but investigating exactly what factors determine that is an exciting question that could have extensive ramifications.

Now you may ask yourself, are these patterns statistically significant? There are many caveats to the conclusions drawn here. Ecology is the result of a complex web of biotic and abiotic factors, and difficult to pull apart, but results speak for themselves. You deploy this type of data analysis by creating and testing new hypotheses in the field, and adjusting appropriately when new data presents itself.

Further analysis might involve applying a clustering algorithm to the geographic data seen above. If there’s a data science expert who knows how they could do this, please get in touch!