Data Science and Decision Making: An Elementary Introduction to Modeling and Optimization

David B. Shmoys, Samuel C. Gutekunst, Frans Schalekamp, and D

Jul 25, 2022

CONTENTS

Temporary cover file for testing.

CHAPTER

ONE

WHAT IS OPERATIONS RESEARCH?

It is hard to give a meaningful definition of operations research. For one, it is hard to say what exactly what is operations research (often referred to simply as OR), in contrast to the the emerging areas of data science, or analytics, both of which have been gaining a lot of attention in recent years. For operations research, one such succinct statement might be:

a quantitative approach to decision-making in which a mathematical model of the problem setting is analyzed so as to provide precise guidance for attaining a desired objective.

However, this definition suffers from an attribute that is not unusual — this definition is based on still other terms that themselves would require a definition. For example, what is a "mathematical model"? What makes an approach "quantitative"? What is meant by "desired objective"? How does a mathematical model get "analyzed"?

Furthermore, there is nothing in any of this explanation that provides some understanding of why operations research is a reasonable name for the discipline. One easy answer is: it is not a good name for the discipline. A somewhat more useful answer is: the roots of the field go back to helping guide "operational" decisions. But what is an "operational" decision?

In the summer of 2020, there were a number of operational decisions that needed to be completely rethought because of the COVID-19 pandemic, starting with, is it "safe" to reopen the campus for in-person education? That led to a million-and-one follow-up questions, all seeking to provide guidance to the more general question, how might we most safely "run" the university if we elect to have in-person education. This is still a very general question, but one might ask a more precisely focused one — if we have target limits on the number of COVID-19 cases that might occur at Cornell throughout the semester, and we can test each student every k days, what is the smallest value of k that provides reasonable assurance of not exceeding those targets. Here we again need to ask, what is "reasonable assurance". One element of building a model for this sort of question relies on the mathematical framework of probability, which captures notions of "likelihood". And partnering with such a probabilistic framework are statistical tools: for example, one might model one mode of transmission by the statement — if two people are within 6' of each other for one hour, there is an α percent chance of a person who is positive for COVID-19, infecting the other — but this gives rise to the associated statistical problem, how can you use historical data to estimate α? Indeed, models to answer this type of question, and are still being used to guide the ongoing decisions that inform the university's response to the pandemic, and have been led by Cornell faculty {\it and students} in operations research.

Although probabilistic and statistical models are an extremely important aspects of operations research, this course will focus on so-called deterministic optimization models (i.e., those without a probabilistic element in their set-up); probability and statistics are (mostly) left to be introduced in the subsequent gateway course in operations research, ENGRD 2700.

Let us start with a very simple example of an optimization problem that was, for those students attending this class in person in Fall 2020, literally staring them in the face. Each student attending in person was in an assigned seat that is at least 6' from anyone else. (To be precise, the center of each occupied seat was at least 6' from the outer edge of any other occupied seat.) The seats are in fixed positions. How do we select in which seats to assign students so that the maximum number of students possible can attend a given lecture? This is an operational question, since its answer is one critical element in the functioning of the university. This course will provide a precise language to state this optimization problem, and provide algorithmic and computational tools to solve problems of this sort - in fact, one later lab exercise will be to attack this very application. OR optimization tools were used heavily in redesigning the Cornell fall course roster, which needed to be completely reworked between the time that the decision was made to reopen with in-person classes in early July and August 26th, when enrollment commenced

Hopefully, the examples above have given you a rough idea about some of the kinds of questions that can be addressed with the tools developed by OR, but these are just a very limited set of examples (however important they were to Cornell over the past few months). One of the exciting aspects of OR is that the applications of this discipline come from many, many different areas — health care, environmental preservation, computer design, transportation logistics, financial instruments, genetics, \ldots, the list goes on and on.

A good next step might be to consider one specific example, and study it more in depth, to help understand the rather hard-to-interpret definition of OR given above.

A colleague who was studying for his PhD in astronomy posed following question. For his dissertation, he was studying a particular set of stars. Every few months, he would be allocated one week of access to a powerful radio telescope. The telescope was programmable, and throughout the week, would focus on one particular star for a given length of time (roughly 10 minutes), acquiring data from the signal from that star, and then would be re-positioned to focus on the next star, and so forth. Since it was a radio telescope, this proceeded for a 24-hour cycle (e.g., it could receive the signal 24/7), when the process would begin anew. The time spent re-positioning was, from his perspective, wasted time. It is easy to imagine if the telescope proceeded through the stars in a completely random order, there might be quite a lot of time wasted in re-positioning. In words, the optimization problem (very roughly) is to maximize the amount of time left to observe the stars under investigation.

One natural question is: how were these decisions being made before? The answer here seemed pretty worrisome — there was a "master list" of all of the known stars in the sky, and this induced an ordering of the relatively few stars in the set that were being studied by this colleague. That determined which star would be observed next.

We want to build a mathematical model to guide this decision-making process. The first element of such a model is to understand what the input to such a model might be - what data do we need in order to provide advice for this colleague? One first idea might be to that the desired input is the positional location of the stars being studied. While this is clearly useful, it might not be the entire story. The goal here is to spend as little time as possible re-positioning the telescope; that is, we need to know how the time to re-position the telescope depends on the positions of the stars observed consecutively. So, for example, if we might move from observing Alpha Centuri to next observing Beta Orionis, we need to know the time that would elapse between completing the first observation, and being ready to start the second, which might be denoted

$$\text{Focus-time[Alpha Centuri, Beta Orionis]}$$

To simplify matters, suppose that there are 100 stars that we wish to observe, and we will call them $1, 2, \ldots, 100$. We denote the set of all 100 stars by $\{1, 2, \ldots, 100\}$ — this is the usual mathematical notation for denoting the set consisting of the integers 1 through 100. In general, we will let n denote the number of stars in our observational set (i.e., $n = 100$). We can use i and j as variables to denote two arbitrary stars in our set — this would be denoted $i \in \{1, 2, \ldots, 100\}$ and $j \in \{1, 2, \ldots, 100\}$, where $i \neq j$. We would then need the data, Focus-time[i,j], for each such pair of stars.

By this point, someone who has studied OR a bit (say, a senior majoring in it at Cornell), but has not seen this telescope observational problem might have the bright idea — "oh, this is just the traveling salesman problem!" And indeed, they would be, more or less, completely right. In words, the traveling salesman problem is often stated as follows: a peddler needs to visit each city exactly once in a given set of cities, starting and ending at home, so as to minimize the total time spent traveling. How should the peddler proceed? In our problem, we have stars, not cities; we have re-positioning times, not travel times, but what we have done is the first conceptual steps in modeling our new application as a traveling salesman problem, which is the next topic in this course.