Multiple data sources model time model

Johnny Kelsey

August 2025

1 Multiple data sources model

How could we model the time taken by a business team having to check internal and external data sources for an address?

Let's assume there are N internal data sources, and M external data sources.

The time taken to check these sources varies.

A stochastic process model is a potentially useful way to model the time taken to check addresses. This approach recognises that the time taken for each check is not a fixed quantity, but is a random variable, which better reflects real-world variability.

2 Building the model

First, we need to define the variables for each type of data source.

- T_{int}: The random variable representing the time taken to check a single internal data source.
- T_{ext}: The random variable representing the time taken to check a single external data source.
- N: The number of internal data sources (constant).
- M: The number of external data sources (constant).
- X: The total time taken to find a valid address.

We assume that the time taken to check each source is an *independent and identically distributed* (i.i.d.) random variable. This means the time to check one source has no effect on the time to check a different data source.

The process of checking for an address is sequential - a team member can only check a single data source at a time. A team member checks a source, and if the address is found and verified, the process stops. Otherwise, they move to the next source.

We assume the team member checks internal sources first, followed by external sources.

The total time, X, is the sum of the times taken to check each source until the address is found.

We have

$$X = \sum_{i=1}^{k} T_{i]} \tag{1}$$

where k is the number of sources checked and T_i is the time taken for the i-th source. Here, k is also a random variable, since we are assuming that the address could be found in any of the data sources.

We model the time taken for each data source check using a probability distribution. An appropriate distribution for this context would be the *exponential distribution.* This is the probability distribution of the distance between events, commonly used to model a process in which events occur continuously and independently, at a constant average rate.

We say that the time to check an internal source, T_{int}, follows an exponential distribution with rate λ_{int}. The average time to check an internal source is $\frac{1}{\lambda_{int}}$.

Similarly, we assume that the time to check an external source, T_{ext}, follows an exponential distribution with rate λ_{ext}. The average time to check an external source is $\frac{1}{\lambda_{ext}}$.

We also introduce probabilities for success at each data source:

- p_{int}: The probability of finding the correct address in any single internal source.
- p_{ext}: The probability of finding the correct address in any single external source.

Recall that we assume the process is the following: the team member checks internal sources first, one by one. If they fail to find a good address after checking all N internal sources, they move on to the M external sources.

The total time X can be broken down into two components:

- Time spent on internal sources: The time taken to check internal sources until a good address is found, or all N sources have been checked.
- Time spent on external sources: The time taken to check external sources, but only if no good address was found in the internal sources.

2.1 Calculating the expected total time

While the total time X is a random variable, we can calculate its expected value, $\mathbb{E}[X]$, which is the average time a team member would spend on this task. The expected time is the sum of the expected times for each step, weighted by the probability of reaching that step.

A fairly simple approach is to think in terms of the probability of success at each stage.

We have:

- Probability of success at source 1: p_{int}.
- Expected time: $p_{int} \times \mathbb{E}[T_{int}]$.
- Probability of success at source 2: $(1 - p_{int}) \times p_{int}$.
- Expected time: $(1 - p_{int}) \times p_{int} \times 2 \times \mathbb{E}[T_{int}]$.
- ... and so on ...

We can simplify and use a more direct approach by considering the states.

- Let P_{Ifail} be the probability of failing to find the address in all N internal data stores. Then we have $(1 - p_{int})^N$.
- Let $\mathbb{E}\left[T_{Itotal}\right]$ be the expected time spent checking all of the internal data sources,
- Let $\mathbb{E}\left[T_{Etotal}\right]$ be the expected time spent checking all of the external data sources.

Let's work through the logic.

2.2 Checking internal data sources

We would like to find a formal way of describing the process of finding a valid address, which is the sum of the time spent checking each internal source, weighted by the probability of the search ending at that source.

This is how we describe the events in the process:

- T_{int}: The random variable representing the time to check a single internal source.
- $\mathbb{E}\left[T_{int}\right]$: The expected time taken to check a single internal data source.
- p_{int}: The probability of finding a good address in any single internal data source.
- N: the total number of internal data sources.

The search process is sequential; it stops as soon as a valid address is found.

- Case 1: Success at the 1st internal source.
  - Probability: p_{int}.
  - Time spent: T_{int}.
  - Expected time spent: $p_{int} \times \mathbb{E}\left[T_{int}\right]$
- Case 2: Failure at the 1st source, but success at the 2nd.
  - Probability: $(1 - p_{int}) \times p_{int}$.

  - Time spent: $2 \times T_{int}$.
  - Expected time spent: $(1 - p_{int}) \times p_{int} \times 2 \times \mathbb{E}\left[T_{int}\right]$

- Case 3: Failure at the 2nd source, but success at the 3rd.
  - Probability: $(1 - p_{int})^2 \times p_{int}$.
  - Time spent: $3 \times T_{int}$.
  - Expected time spent: $(1 - p_{int})^2 \times p_{int} \times 3 \times \mathbb{E}\left[T_{int}\right]$

This pattern continues for all of our N internal data sources. The total expected time for the internal search is the sum of the expected times for all possible successful outcomes.

Putting this all together into an equation for the expected time taken for each outcome, we have:

$$\mathbb{E}\left[T_{Totalint}\right] = \sum_{i=1}^{N} (1 - p_{int})^{i-1} \times p_{int} \times i \times \mathbb{E}\left[T_{int}\right] \tag{2}$$

This equation represents the sum of the probabilities of finding the address at each specific source, multiplied by the time it would take to reach that source.

2.3 Checking external data sources

We assume that the team member has exhausted all the internal data sources - the address might be missing, or it might be incomplete, or they might be trying to verify or validate the data.

The key insight is that a team member only starts checking external sources *if and only if* they fail to find a valid address in all of the internal sources. This makes the time spent on external sources a conditional event.

- P_{Ifail} is the probability that the internal search fails. This is the probability of the first internal data source failing, and the second failing, ..., and the N-th source failing.
- So we have $P_{Ifail} = (1 - p_{int})^N$

So the expected total time spent on internal sources. $\mathbb{E}[T_{Etotal}]$, can be expressed informally as

$$\mathbb{E}[T_{Etotal}] = P_{Ifail} \times \mathbb{E}[P_{Text}|P_{Ifail}] \tag{3}$$

where P_{Ifail} is the probability of being unable to confirm the data using internal sources, and P_{Text} is the probability of a certain time taken to locate the address in external sources. So we have a conditional expectation, such that the total expected value of an event is the probability of its condition occurring, multiplied by the expected value of the event given that condition. Note that we have $\mathbb{E}[P_{Text}|P_{Ifail}]$, which is the expected time to find an address among the external data sources, given that all internal sources have been exhaustively checked, and failed to find the address. The model assumes that the internal and external searches are independent processes, so this simply becomes the expected time of the external search process itself.

Breaking this down further, the expected time taken to find an address in a sequence of M external sources is the sum of the expected times for each of the possible outcomes. Let T_j be the time taken for the j-th external source.

- Case 1: Success at 1st external source.
  - Probability: p_{ext}
  - Time: $1 \times \mathbb{E}\left[T_{ext}\right]$.
- Case 2: Success at 2nd external source.
  - Probability: $(1 - p_{ext}) \times p_{ext}$.
  - Time: $2 \times \mathbb{E}\left[T_{ext}\right]$.
- $\cdots$
- Case M: Success at the M-th external source.
  - Probability: $(1 - p_{ext})^{M-1} \times p_{ext}$
  - $M \times \mathbb{E}\left[T_{ext}\right]$
- Case $M + 1$: Failure to find the address after checking all M sources.
  - Probability: $(1 - p_{ext})^{M}$.
  - Time: $M \times \mathbb{E}\left[T_{ext}\right]$

So the total expected time is the sum of the expected time for each of the above cases:

$$\mathbb{E}\left[Text\right] = \sum_{j=1}^{M} (1 - p_{ext})^{j-1} \times p_{ext} \times j \times \mathbb{E}\left[T_{ext}\right] + (1 - p_{ext})^{M} \times M \times \mathbb{E}\left[T_{ext}\right] \quad (4)$$

This formula is the standard way to calculate the expected number of trials in a limited geometric distribution (recall that we have a finite number of data sources), We get those probabilities and multiply them by the expected time per trial.

3 Simulation

The model above can be translated into code, and used to simulate the process of an individual team member looking for an address over a number of different trial runs.

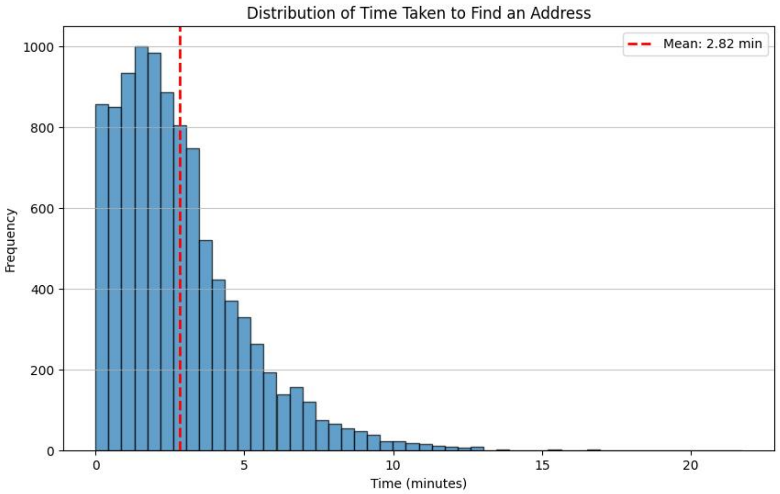

Figure 1: Time taken to check multiple data sources.

The code allows us to see the results of running the model given different parameters. In order to run the code, we need to define a few parameters. We define the following:

- number of internal data sources: 3
- number of external data sources: 5
- p_{int}: the probability of finding the correct address in any single internal source: 0.1.
- p_{ext}: the probability of finding the correct address in any single external source: 0.1.
- average time taken to check internal data sources: 0.5 (minutes).
- average time taken to check external data sources: 2.82 (minutes).

The model was run for 10000 trials. Figure (1) shows the results.