COVID-19 Campus Simulation
Models Description

Dhaval Adjodah, Jim DiCarlo, Katherine Fairchild, Ruijie He, Josh Joseph,
Kyle Keane, Nicholas Roy, Erik Vogan

November 5, 2021

Summary

We describe two complementary COVID-19 models used at MIT in this document:

- The MIT Situational Awareness (SA) system used to ingest live data streams and to predict if short-term (hours) building occupancy, inflows, and outflows remain at safe levels. Section 1 describes the design requirements and mathematical formalism behind the SA system.
- The COVID-19 Risk Model (RM) used for longer term (weeks) prediction of building usage. Section 2 describes the design requirements and mathematical formalism behind the RM system.

Note that many parts of these systems were designed around March 2020. This was before information about vaccines, variants, etc. were known. A sanitized version of the codebase for the RM and SA systems can be found in our Github repository. For privacy and security reasons we have chosen to not open source the complete code base that is used in the operational system. For example, we have include synthetic data and removed code that handles integration with MIT's IT infrastructure.

Contents

1 Situational Awareness

1.1 Overview

In contrast to the Risk Model system described in section 2, which we ran weekly to simulate a number of policy actions in parallel, we run the situational awareness (SA) system hourly. SA provides decision makers with hourly predictions of building usage statistics, and alerts them of any buildings where occupancy has reached a high thresholds of density.

1.2 Overarching assumptions of this system

1. We are not estimating health risk. Instead we estimate building density (occupancy per usable square footage) which is used by decision makers as a proxy for health risk (the denser a building is, the more likely infections will spread). Similarly, we provide building inflow and outflow as a way to minimize crowding at points of entry/exit.

2. We assume that building usage threshold alerts have been provided to us based on decision makers' assumptions of virus transmission and floor plans.

3. Decision makers need high frequency estimates of **future** building usage in order to assign building access, plan events, and generally control building density levels.

4. Decision makers need high frequency estimates of **past** building usage in order to verify and build intuition as to how their past decision impacted building usage.

5. We expect that people will swipe into a different building that the one they plan to go to. e.g. they will swipe into an entry building and walk through several connected buildings to reach their office.

6. We expect that building usage will be different for different days depending on whether they are regular week days, weekends, holidays, institute holidays, etc.

7. It is better to overestimate building usage than to underestimate it due to the higher health cost of underestimation.

1.3 High-level design decisions

1. We assume that only people with active assignments (badge access) to certain buildings are allowed into these buildings.

2. Our smallest unit of geographic estimation are buildings, not floors or rooms.

3. We do not estimate past inflow into building and instead data as-is.

4. We assume that people only come into campus for one "trip", i.e. there are no repeated visits. Although this is not always correct (people leave and come back), model checking has shown that modeling repeated visits leads to huge overestimates.

5. We use anonymized data to estimate building usage duration (how long somebody stays in a building) and building-to-building transition matrices. We only include transitions that occur within a 15min window. We do not retain a copy of the source data and it is regularly purged as per data retention policies.

6. Given that computing duration and transition parameter distributions takes a long time, we only compute it once a week.

7. We assume that using one month of past data is enough history to estimate parameters over. Model checking shows that less time than 1 month leads to high noise. Longer time history leads to missing campus important calendar changes (e.g. holidays skew our estimates).

8. We expect that people stay overnight in their building and therefore model overnight trajectories.

9. We inject building inflows, outflows and occupancy estimated from class schedule on top of estimated values. The amount of this injection decreases linearly as number of people on campus estimated from data increases.

1.4 Definitions

1.4.1 Time

Let time t be an hourly timestamp (e.g. 2020-04-01 19:00). It is not a random variable. t_0 is the time of the beginning of the simulation, which is, in implementation, 1 month earlier than the current hour where the simulation is being run.

For estimating distributions, we operate over 'hour of the week' which is the number of hours since Monday 12am. It reaches 167 (zero-indexed) on the Sunday 11pm of that week. We define the function $h \in \{0, ..., 167\}$ where $h(t)$ provides the mapping from a timestamp t to h. For example, 2020-04-01 19:00 maps to 67.

1.4.2 Building location

Building locations, l, are defined as the building name under investigation, e.g. 'NW14'. It is not a random variable.

1.4.3 Occupancy, inflow and outflow

Our goal is to estimate the random variable occupancy $O_{[t,t+1),l}$ in a building l during the interval $[t,\ t+1)$.

Similarly $I_{[t,t+1),l}$ is the random variable inflow and $X_{[t,t+1),l}$ is the random variable outflow.

1.5 Mathematical Derivation

1.5.1 Inflow model

We start with estimating inflow, which is the number of arrivals:

$$I_{[t,t+1),l} \sim \mathbb{I}(\theta(h(t), l)) \tag{1}$$

The parameters $\theta(h(t))$ of the distribution $\mathbb{I}$ are estimated using Maximum-Likelihood Estimation (MLE). For example, we can use arrivals over the past Tuesdays between 12-1pm to estimate a distribution of arrivals during hour 36. In implementation, we estimate a Normal distribution's mean $\mu_{[h(t),h(t)+1),l}$ and standard deviation $\sigma_{[h(t),h(t)+1),l}$

For example, if $I_{[t,t+1),NW14} = 5$, where t = 2020-01-12 09:00, this means there were 5 arrivals at this time.

We further define $i_{[t,t+1),l} \in [0, ..., I_{[t,t+1),l}]$ to be a sequence of counters to be used in sums in the next section. In the previous example where $I_{[t,t+1),NW14} = 5$, this leads us to $i_{[t,t+1),NW14} \in [0, 1, 2, 3, 4]$.

In our implementation, I is more finely estimated, for example, based on whether the day of the week was a holiday. For simplicity, we do not present the full picture here.

1.5.2 Duration model

For each time t, for each building l, and for each arrival $i_{[t,t+1),l}$ (as different people arriving at the same time might stay different amounts of time), we define its duration:

$$D_{i_{[t,t+1),l},[t,t+1),l} \sim \mathbb{D}(\zeta(h(t), l) \tag{2}$$

In our implementation, $\mathbb{D}$ is an empirical (categorical)x distribution with ζ estimated using MLE from data.

We further define γ as the sequence of timestamps over which each arrival $i_{[t,t+1),l}$ was present in a building l: $\gamma_{i_{[t,t+1),l},l} \in [t, t+1, ..., t + D_{i_{[t,t+1),l},[t,t+1),l}]$. E.g. if somebody came in at 2020-10-12 09:00 and stayed for 3 hours, their sequence would be [2020-10-12 09:00, 2020-10-12 10:00, 2020-10-12 11:00].

1.5.3 Occupancy model

Given the above models and definitions,

$$O_{[k,k+1),l} = \sum_{t=t_0}^{k} \sum_{i_{[t,t+1),l}} \sum_{\gamma_{i_{[t,t+1),l},l}} \mathbb{1}(k \leq \gamma_{i_{[t,t+1),l},l} < k+1) \tag{3}$$

where $\mathbb{1}$ is the binary indicator variable, k is a timestamp (similar to t) over which occupancy is being evaluated.

As an intuitive explanation, the inner sum is summing occupancy for a particular time interval $[k, k+1)$ over which an arrival i_t is still in this building, summed (middle sum) over all arrivals, summed (outermost sum) over all timestamps up to the beginning of the simulation (e.g. a person that came in at the beginning of the simulation with a very large duration could still be in this building).

One assumption of the above model is that a person coming into a building only stays within this building. Our data shows that this is not always the case, and therefore one needs to estimate transition between arrival buildings (where people swipe in) and 'occupancy' buildings (where their occupancy should be attributed). We do so using the stochastic transition function Ψ described next.

1.5.4 Transition model

The stochastic function $\mathcal{T}$ simulates transitions of people from the building b they swipe into to their work building b' they eventually end up in. Occupancy is only assigned to b', while inflow and outflow are assigned o both buildings, in addition to the buildings in the walking path $\mathbf{p}$ (a vector of buildings) between the two buildings. The stochastic function $\mathcal{T}$ is implemented as algorithm 1.

To use $\mathcal{T}$, we need to define two new distributions. Using data, we estimate, $\mathbb{T}$, a distribution of transitions between buildings. Using $\mathbb{T}$, given a timestamp t and a starting building l, we can sample a transited building $l'_{[t,t+1),l}$:

$$l'_{[t,t+1),l} \sim \mathbb{T}(\kappa(h(t), l)) \tag{4}$$

Similarly, we compute assignment distribution $\mathbb{A}$ which allows us to sample a building $l''_{[t,t+1),l}$ that a person is assigned to given their swipe building. Specifically, what we are trying to model here is based on the fact that, behind the scenes, each person i is 'assigned' to the buildings they need to work in (e.g. their office, labs, machine shops, etc.). However, they often need access to other buildings to enter through (e.g. because their work building is not on a public street, so they need to access it through corridors belonging to other buildings).

$$l''_{[t,t+1),l} \sim \mathbb{A}(\xi(h(t), l)) \tag{5}$$

κ and ξ are estimated using MLE from data.

Algorithm 1 Stochastic building-to-building transition function $\mathcal{T}$

Input: Swipe (entry) building b, timestamp t, transition distribution $\mathbb{T}$, assignment distribution $\mathbb{A}$.
Output: Work building b', walking path $\mathbf{p}$.
Hyperparameters: Maximum number of attempted transitions N_T, Maximum number of attempted assignments N_A.
for assignment trial $j = 0, \ldots, N_A$ **do**
  current building: $b_{current} = b$
  walking path: $\mathbf{p} = \varnothing$
  Sample work building: $b^{trial}_{[t,t+1),b} \sim \mathbb{A}(\xi(h(t), b))$
  if $b^{trial}_{[t,t+1),b} == b$ **then**
    `# end if sampled trial work building is your starting building`
    return: $b^{trial}_{[t,t+1),b}$, $\mathbf{p}$
  for transition trial $k = 0, \ldots, N_T$ **do**
    sample transition building: $b^{transition}_{[t,t+1),b_{current}} \sim \mathbb{T}(\kappa(h(t), b_{current}))$
    $\mathbf{p} := \mathbf{p} \bigcup \{b^{transition}_{[t,t+1),b_{current}}\}$
    if $b^{transition}_{[t,t+1),b_{current}} == b^{trial}_{[t,t+1),b}$ **then**
      `# end if sampled transition building is your trial work building`
      return: $b^{trial}_{[t,t+1),b}$, $\mathbf{p}$
    else
      $b_{current} = b^{transition}_{[t,t+1),b_{current}}$
      `# continues inner for loop`
`# return starting building if no work building was found`
return: b, $\varnothing$

Based on the definitions in section 1.4 and the derivation in section 1.5, we built an automated system that simulates hourly building occupancy and sends alerts if any building's usage is above a certain safety threshold.

The code and readme documents the implementation of this system.

1.6 Model and implementation checking

We implement a number of ways to check our implementation:

- Implementation check: we verify that the conservation of people is respected during our simulation interval per building i.e. the total number of people entering a building is the same as the total number of people leaving.
- Model check: we implement a predictive check to compare simulated past values to manually collected values. For example, we had people stand outside all doors of buildings and count the number of people coming in and out over a time-interval, and this value was compared to

our simulated values as a predictive check. This is implemented in the `create_ground_truth_comparison` function.

2 COVID-19 Risk Model

2.1 Overview

In this section, we describe the COVID-19 Risk Model system (which is separate from the Situational Awareness system described in section 1). The goal of this model is to provide long term (weeks) prediction of building usage and risk conditional on policy actions. We first define what we mean by "COVID-19 risk", and then define various models in the pipeline.

2.2 System assumptions

Note that the following system assumptions were made in March 2020. This was before information about vaccines, variants, etc. were known.

1. We will have not achieved herd immunity or have a vaccine for SARS-CoV-2 until at least 2021.
2. Coordinated government guidance will need to be augmented by critical decisions informed by local environmental considerations. In other words, institutions like MIT will need to make many informed decisions to safeguard their populations.
3. We believe approaches such as contact tracing and exposure alerting apps are important tools but these needs were being addressed by others. Forecasting risk well for a small, target population (e.g., MIT's community) is extremely hard and could prevent a great deal of harm.
4. Immunity and infection (i.e. person state transitions) will be treated as a single date when that state transition takes place (in a binary fashion).
5. We will not be able to estimate everything perfectly. However, using posterior predictive checks, we can estimate our prediction error.

2.3 High-level design decisions

1. The model will not attempt to combine different types of risk (e.g., health and economic).
2. Our models, parameters, and data are going to be inaccurate. Therefore, modeling uncertainty well is crucial.
3. We do not believe all the important benefits and harms can be quantified or well-modeled, and believe the sort of health model provided here will just be one of many factors that the institute's administration will use to make final policy decisions.
4. The model must be able to show the effect of interventions on different population demographics.

5. We think of risk as the harm (e.g., sickness, death, job loss, community cohesion loss) to a population.

6. The distribution over harm is an important model output as we do not assume we know the most appropriate statistic of the distribution useful for making decisions (e.g., expected harm, 95-percentile harm).

7. While overall risk is composed of various subcategories of COVID-19 related risk (i.e.: health risk, economic risk, and social risk), here, we focus only on health risk.

8. Inclusion of an individual's mobility will allow us to model movement-related interventions impact on SARS-CoV-2 prevalence.

9. The risk model will be composed of a handful of sub-models. We do not assume we know the structure or parameters of the sub-models and allow for a wide range of possible future sub-models to be used.

2.4 Definitions

2.4.1 Time

Define T^V be a random variable representing the end date of our simulation. E.g. this could be the immunity date at which we no longer need be concerned with any additional COVID-19-related risk (e.g., either when a vaccine is available and widely administered and/or herd immunity is achieved).

2.4.2 Individuals

Define i be a random variable representing an individual i drawn from some population I. For example, I may be all residents of the greater Boston area or all MIT employees or MIT undergraduates who attend class in building 32.

Define $\texttt{new-infection}(i, t)$ to be a random indicator variable denoting if i has become infected with SARS-CoV-2 on day t. Note $\texttt{new-infection}(i, t)$ is true for at most one value of t and false otherwise.

Define $\texttt{becomes-infected}(i, t_0, t_1)$ to be a random indicator variable denoting if i becomes infected with SARS-CoV-2 on day between t_0 and t_1.

Define $\texttt{demographics}(i)$ to be the COVID-19 demographics of i.

2.4.3 Actions

Define $a_{\tau:T^V}$ to be the series of actions taken between days τ and T^V. For example, $a_t \in [0, 100]$ could be the percentage of people we allow to return to campus, or to a specific building.

2.4.4 Health Risk

Define the possible health-related harms from COVID-19 as $H = \{\text{sickness, hospitalization, needing ICU care, death}\}$ and $\texttt{experience-harm}(i, h)$ to a random indicator variable representing individual, i, experiencing harm, h.

Define the risk to the population I from actions $a_{t_0:t_1}$ taken between days t_0 and t_1 to be the set of distributions over the total harms experienced across the population:

$$R_{health}(I, a_{t_0:t_1}) \coloneqq \left\{ \left[h, P\left(\sum_{i \in I} \texttt{experience-harm}(i, h) \middle| a_{t_0:t_1} \right) \right] \right\}_{h \in H} \tag{6}$$

$$= \left\{ \left[h, \sum_{i \in I} P\left(\texttt{experience-harm}(i, h) \middle| a_{t_0:t_1} \right) \right] \right\}_{h \in H} \tag{7}$$

Equation 7 is the output of the COVID-19 Risk model (as a distribution) that we provide to decision makers in the form of summary statistics (e.g. median, confidence intervals).

To calculate this health risk, we make a the independence assumptions below, and then build different modules that use Monte-Carlo samples to estimate this risk.

In actuality, we are less concerned with absolute health risk than we are with additional health risk resulting from taking a set of actions. So we define a set of baseline actions (e.g., no-ops) $a^{baseline}_{\tau:\infty}$ and compute the relative risk in taking actions $\tilde{a}_{\tau:\infty}$:

$$\Delta R_{health}(I, \tilde{a}_{\tau:\infty}, a^{baseline}_{\tau:\infty}) \coloneqq R_{health}(I, \tilde{a}_{\tau:\infty}) - R_{health}(I, a^{baseline}_{\tau:\infty})$$

$$= \left\{ \left[h, \sum_{i \in I} P\left(\texttt{experience-harm}(i, h) \middle| \tilde{a}_{\tau:\infty} \right) - P\left(\texttt{experience-harm}(i, h) \middle| a^{baseline}_{\tau:\infty} \right) \right] \right\}_{h \in H} \tag{8}$$

2.5 Mathematical Derivation

In order to estimate the population risk, we first derive an approximation (given independence conditions) of equation 7.

2.5.1 Health risk model for individual i

The probability of harm h to an individual i when actions $a_{\tau:\infty}$ are taken starting on day τ is:

$$P(\texttt{experience-harm}(i, h) | a_{\tau:\infty}) \tag{9}$$

As per the law of total probability,

$$= \sum_{T^V} P(\texttt{experience-harm}(i,h), T^V | a_{\tau:T^V}) \tag{10}$$

Based on the definition of conditional probability,

$$= \sum_{T^V} P(T^V) P(\texttt{experience-harm}(i,h) | T^V, a_{\tau:T^V}) \tag{11}$$

Because no harm is experienced without infection,

$$= \sum_{T^V} P(T^V) P(\texttt{experience-harm}(i,h), \texttt{becomes-infected}(i,\tau,T^V) = True | T^V, a_{\tau:T^V}) \tag{12}$$

Again, based on the definition of conditional probability,

$$= \sum_{T^V} P(T^V) P(\texttt{experience-harm}(i,h) | \texttt{becomes-infected}(i,\tau,T^V) = True) \times$$
$$P(\texttt{becomes-infected}(i,\tau,T^V) = True | T^V, a_{\tau:T^V}) \tag{13}$$

Because you can only be infected once (this assumption was pre-multi variants), `becomes-infected` is the result of past `new-infection`,

$$= \sum_{T^V} P(T^V) P(\texttt{experience-harm}(i,h) | \texttt{becomes-infected}(i,\tau,T^V) = True) \times \sum_{t=\tau}^{T^V} P(\texttt{new-infection}(i,t) = True | a_{\tau:t}) \tag{14}$$

Now we can introduce `mobility` as:

$$P(\texttt{new-infection}(i,t) = True | a_{\tau:t}) \tag{15}$$

By the law of total probability,

$$= \sum_{\texttt{mobility}(i,t,a_{\tau:t})} P(\texttt{new-infection}(i,t) = True, \texttt{mobility}(i,t,a_{\tau:t}) | a_{\tau:t}) \tag{16}$$

Based on the definition of conditional probability,

$$= \sum_{\texttt{mobility}(i,t,a_{\tau:t})} P(\texttt{mobility}(i,t,a_{\tau:t}) | a_{\tau:t}) P(\texttt{new-infection}(i,t) = True | \texttt{mobility}(i,t,a_{\tau:t}), a_{\tau:t}) \tag{17}$$

Assuming that action $a_{\tau:t}$) affects new infection only via mobility i.e. mobility is a sufficient statistic of action,

$$= \sum_{\texttt{mobility}(i,t,a_{\tau:t})} P(\texttt{mobility}(i,t,a_{\tau:t})|a_{\tau:t}) P(\texttt{new-infection}(i,t) = True|\texttt{mobility}(i,t,a_{\tau:t})) \tag{18}$$

2.5.2 Mobility and Infection Models

Let us assume that infection, $P(\texttt{new-infection}(i,t) = True|\texttt{mobility}(i,t,a_{\tau:t})))$, is a function of three statistics:

1. $\texttt{interactions}(i,t,a_{\tau:t})$: the number of person-to-person interactions i has on day t
2. $\texttt{shared-space}(i,t,a_{\tau:t})$: the number of people who shared the same physical space as i on day t, albeit at different times
3. $\texttt{prevalence}(i,t)$: the prevalence of SARS-CoV-2 on day t in the local population i interacts or shares space with (for now, let's interpret that to mean the prevalence in the zip codes i inhabits)

If we assume these two effects contribute independently, we get:

$$P(\texttt{mobility}(i,t,a_{\tau:t})|i,a_{\tau:t}) = P(\texttt{interactions}(i,t,a_{\tau:t}), \texttt{shared-space}(i,t,a_{\tau:t})|i,a_{\tau:t})$$

Assuming conditional independence,

$$= P(\texttt{interactions}(i,t,a_{\tau:t})|i,a_{\tau:t}) P(\texttt{shared-space}(i,t,a_{\tau:t})|i,a_{\tau:t}) \tag{19}$$

To compute these, it's useful to decompose them into *when* during a day these interactions and shared spaces happen:

1. i's commute to campus
2. i's walk across campus to their building
3. inside of the building which i works
4. during i's non-work activities work around campus

During each of these times we can create separate models of interactions and shared spaces. Historical summaries of location occupancy provide a reasonable worst case model (e.g., people during a pandemic are more likely to avoid people than prior to a pandemic, so historical data is likely to overestimate these quantities).

Also note `interactions` and `shared-space` are the first effect we will see from many of the actions we take.

The second part of equation 18 requires us to estimate:

$$P(\texttt{new-infection}(i,t) = True|\texttt{mobility}(i,t,a_{\tau:t}))$$

By the law of total probability,

$$= \sum_{\texttt{prevalence}(i,t)} P(\texttt{new-infection}(i,t) = True, \texttt{prevalence}(i,t)|\texttt{mobility}(i,t,a_{\tau:t})) \tag{20}$$

By the definition of conditional probability,

$$= \sum_{\texttt{prevalence}(i,t)} P(\texttt{new-infection}(i,t) = True|\texttt{prevalence}(i,t),\texttt{mobility}(i,t,a_{\tau:t}))\times P(\texttt{prevalence}(i,t)|\texttt{mobility}(i,t,a_{\tau:t})) \tag{21}$$

Assuming that the prevalence (at the region level, e.g. zipcode) does not depend significantly on mobility of a single individual i,

$$= \sum_{\texttt{prevalence}(i,t)} P(\texttt{new-infection}(i,t) = True|\texttt{prevalence}(i,t),\texttt{mobility}(i,t,a_{\tau:t}))\times P(\texttt{prevalence}(i,t)) \tag{22}$$

Substituting equation 19 regarding the independence of mobility components:

$$= \sum_{\texttt{prevalence}(i,t)} \Big[P(\texttt{new-infection}(i,t) = True|\texttt{prevalence}(i,t),\texttt{interactions}(i,t,a_{\tau:t})) + P(\texttt{new-infection}(i,t) = True|\texttt{prevalence}(i,t),\texttt{shared-space}(i,t,a_{\tau:t}))\Big]\times P(\texttt{prevalence}(i,t)) \tag{23}$$

2.5.3 Independence assumptions

As a brief summary, these are the independence assumptions assumed for the above derivations:

1. $\texttt{experience-harm}(i,h)|a_{t_0:t_1} \perp\!\!\!\perp \texttt{experience-harm}(j,h)|a_{t_0:t_1}$, for $i \neq j$
2. $T^V \perp\!\!\!\perp a_{\tau:T^V}$
3. $\texttt{experience-harm}(i,h)|\texttt{becomes-infected}(i,\tau,T^V) \perp\!\!\!\perp T^V, a_{\tau:T^V}$
4. $\texttt{new-infection}(i,t)|\texttt{mobility}(i,t,a_{\tau:t}) \perp\!\!\!\perp a_{\tau:t}$
5. $\texttt{prevalence}(i,t) \perp\!\!\!\perp \texttt{mobility}(i,t,a_{\tau:t})$
6. $\texttt{new-infection}(i,t)|\texttt{prevalence}(i,t),\texttt{interactions}(i,t,a_{\tau:t}) \perp\!\!\!\perp \texttt{new-infection}(i,t)|\texttt{prevalence}(i,t),\texttt{shared-space}(i,t,a_{\tau:t}))$

2.5.4 Estimating the models we'll need

There are eight total sub-models that we'll need to compute $P(\texttt{experience-harm}(i,h)|i,a_{\tau:\infty})$:

1. $P(\texttt{experience-harm}(i,h)|\texttt{becomes-infected}(i,t_0,t_1)=True,\texttt{demographics}(i))$
2. $P(\texttt{prevalence}(\texttt{region}(i),t)|\texttt{gov-actions}(\texttt{region}(i),t))$
3. $P(\texttt{new-infection}(i,t)=True|\texttt{prevalence}(i,t),\texttt{interactions}(i,t,a_{\tau:t}))$
4. $P(\texttt{new-infection}(i,t)=True|\texttt{prevalence}(i,t),\texttt{shared-space}(i,t,a_{\tau:t}))$
5. $P(\texttt{interactions}(i,t,a_{\tau:t})|i,a_{\tau:t})$
6. $P(\texttt{shared-space}(i,t,a_{\tau:t})|i,a_{\tau:t})$
7. $P(I)$, distribution of demographic of individuals.
8. $P(\texttt{gov-actions}(\texttt{region}(i),t))$

2.6 System structure

Given the previous derivation, we built a system with models approximated using a sampling approach.

A variety of considerations prevent us from releasing the system in current daily use, including privacy and safety considerations and infeasibility of separating model code from campus infrastructure. Here, instead, we present a simplified structure that has been sanitized for public release, as it may benefit others implementing future similar systems.

As represented in figure 1, each model is a block and generates probability samples which are ingested by the next model, with the end goal of producing health risk samples which approximate the population risk.

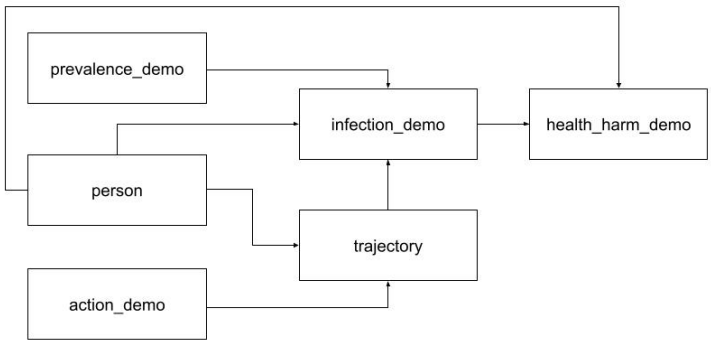

Figure 1: The MCRS Model Structure

This is one of many possible approaches to model health risk. Here, we detail our particular model definitions and sampling methods. However, one could easily create a new type of model, e.g. a 'vaccination' model that simulates vaccination campaigns and its effect or the other models.

As a brief summary, this system models the flow of people (each defined as a `person` sample with demographic, building assignments, and commuting attributes) walking through buildings (`trajectory` samples of people transiting between buildings each hour as a stochastic function of building openings, themselves samples of the `action` model), getting infected (`infection` samples are a stochastic function of local COVID-19 `prevalence` samples per zipcode, number of contacts based on `trajectory` samples, infection susceptibility based on `person` demographics) and the progress of these infections into `healthharm` samples.

Please see the code for details about implementation.

2.7 Managing Uncertainty

There are four main types of uncertainty we have to cope with:

1. Inherent randomness (e.g., the model is probabilistic)
2. Uncertainty due to approximate inference (from using a sampling method)
3. Uncertainty due to incorrect parameters
4. Uncertainty due to incorrect model structure

2.7.1 Inherent randomness

Inherent randomness is something we wish to display explicitly to the user. To compute this, we use a bootstrap estimate of the distribution over the statistic of interest.

For an example of the bootstrap estimation procedure, if we want to show the distribution over the total number of deaths for ages 10-20 and 1,000 runs of MC simulations, we would sample, with replacement, 1,000 runs from our actual samples and compute the total number of deaths for ages 10-20. After repeating this process between 100 and 10,000 times we will obtain the distribution over total deaths for ages 10-20.

2.7.2 Uncertainty due to approximate inference

We can easily quantify and remedy uncertainty due to approximate inference by running a variety of sample sizes, computing their error bars, and increasing or decreasing the sample size to achieve the desired level of error.

2.7.3 Uncertainty due to incorrect parameters and model structure

We provide the `observations_and_model_distributions` analysis module in our code which allows for comparison between simulated variables (building occupancy predictions) communicated to decision makers and observed values. This allows for model validation and improvement in terms of the parameters and model structure used.