A RATIONAL AGENT MODEL FOR THE SPATIAL ACCESSIBILITY OF PRIMARY CARE

JAMES SAXON, DANIEL SNOW

Harris School of Public Policy and Center for Spatial Data Science, University of Chicago

Abstract. Accurate modeling of the spatial accessibility of healthcare is critical to measuring and responding to physician shortages. We develop a new model in which patients choose the primary care location that minimizes their combined accessibility and availability costs. This model improves on existing access frameworks by endogenizing the trade-off between travel times and congestion at the point of care. It allows for patients to seek care from their home or workplace, and can incorporate multiple travel modes. Our open-sourced implementation scales efficiently to large areas and fine spatial granularity. Using distributed computing, we calculate travel times for this model at the Census tract level for the entire United States, and we also make this resource available. With this model and data, we evaluate spatial access costs for primary healthcare and analyze the implications for rural areas. We compare the results to those from existing primary care access models.

1. Introduction

Measurement of the spatial accessibility of primary healthcare is a classic problem in geographic analysis. It is an important determinant of whether or not populations receive care that they need. It currently has acute policy relevance: rural America faces a dramatic shortage of physicians. Where are these shortages, and what drives them?

- Why primary care matters.
- Components of access.
- What we focus on: spatial accessibility (non-monetary, non-SES, etc.)

This paper presents two methodological contributions to this measurement. The first is in modeling: an intuitive and efficient framework offering several conceptual advantages with respect to traditional floating catchment methods. In our framework, which we call the Rational Agent Access Model (RAAM), patients seek care at the physician office with the lowest combined cost of travel time plus congestion at the point of care. This principal, which privileges competition between locations for care, follows that developed by Serban, et al. It endogenizes the trade-off between travel times and office congestion and introduces feedback between patients’ decisions. This realistic, competitive response to congestion has not been available in floating catchment methods. An efficient optimization iterates over patient locations, and terminates when all patients at each residential location experience the same costs – when no improvements are available. This reveals the “cost of care for patients.

RAAM readily incorporates many of the conceptual refinements common in floating catchment methods. We show how changing willingness to travel affects results. Several recent papers explore the impact of multi-modal travel, and we reproduce this functionality as well. Tipping our hats to extensive literatures on activity spaces and time geography that frame human activity with respect to two poles – home and workplace – we allow patients to seek care starting from either origin. This mechanism is, to our knowledge, totally new to the medical access literature. Finally, we present strategies for contrasting modelled results with data. Taken together, these modifications offer a raft of realistic systematics.

The second contribution of our paper is technical. A central challenge for access measurements is the calculation of travel time matrices. We present strategies for calculating large-scale matrices for both driving and public transportation, using open source tools and distributed computing. We provide the outputs – tract-level, national scope matrices – for other researchers. This is particularly important,
given the high and rising costs of commercial travel time (isochrone) providers. These matrices are needed not only for research on healthcare accessibility but for an entire class of problems touching on routing, accessibility, and resource allocation: deliveries, park access, and commercial site placement, for example.

In the present context of the accessibility of rural primary care, this travel time calculation is itself an important advance. It affords our measurement unprecedented scope and granularity.\footnote{We base this statement on the total number of individual origin locations considered and the complexity of the road network. We consider every Census tract in the United States – upwards of 70,000 cells. Others have preceded us with finer granularity or national scale. For example, Li et al perform a county-level calculation for the continental United States. McGrail and Humphreys calculate access to primary care in Australia, using cells with just a few hundred people in rural areas. But the Australian road network is far less complex and the total number of origins is less than a third of ours.} Seeing the country at a glance offers an important and immediate perspective: there is enormous heterogeneity in accessibility, across rural America. The rural shortage is not, perhaps, so simply rural. Before concluding, we therefore explore the variability in rural access through several simple regressions, contrasting the explanatory power of density itself with that of other socioeconomic factors.

2. Modeling Accessibility

2.1. Current Methods and their Limitations. This project stands in a long line of models for healthcare access. We review these briefly before presenting our model and contrasting it with its predecessors. We then describe simple and meaningful modifications of our approach, which we evaluate in subsequent Sections.

2.1.1. The Patient to Provider Ratio. The simplest measure of physician accessibility is the provider to provider ratio (PPR): the average number of patients cared for by a physician. The PPR has a venerable history, and has long been used by the Health Resources and Services Administration (HRSA) for designating Health Provider Shortage Areas (HPSAs). This ratio may be adjusted to account for patient needs or physician capacity.\footnote{We base this statement on the total number of individual origin locations considered and the complexity of the road network. We consider every Census tract in the United States – upwards of 70,000 cells. Others have preceded us with finer granularity or national scale. For example, Li et al perform a county-level calculation for the continental United States. McGrail and Humphreys calculate access to primary care in Australia, using cells with just a few hundred people in rural areas. But the Australian road network is far less complex and the total number of origins is less than a third of ours.} Despite its Federal imprimatur and the appeal of its simplicity, the PPR has major shortcomings. It is unrealistic for small area estimates, because area boundaries are usually permeable: patient populations are not fixed but will instead respond to high or low availability to equilibrate demand. The question of “how many people per doctor” is a good one, but fine partitions of population are suspect since patients are not constrained by these boundaries.

One strategy to address this issue, adopted by The Dartmouth Institute (TDI), is to define regions that better encapsulate primary care utilization patterns. TDI calls these regions Primary Care Service Areas (PCSAs).\footnote{We base this statement on the total number of individual origin locations considered and the complexity of the road network. We consider every Census tract in the United States – upwards of 70,000 cells. Others have preceded us with finer granularity or national scale. For example, Li et al perform a county-level calculation for the continental United States. McGrail and Humphreys calculate access to primary care in Australia, using cells with just a few hundred people in rural areas. But the Australian road network is far less complex and the total number of origins is less than a third of ours.} Though the PCSAs remain permeable, TDI also evaluates “re-allocated” patient demand, using claims data to identify patients’ residences and unwind the permeability issue that arises inevitably from patients’ commutes to care. Yet this still does not fully solve the problem: allocated visits represent realized (rather than potential) access. This realized access may have been costly for patients to capture, and so the original accessibility is still not measured.

2.1.2. Floating Catchment Methods. The natural response is to shift all boundaries away from patients – to define a Floating Catchment Area (FCA) around each patient and define the PPR within that floating area.\footnote{We base this statement on the total number of individual origin locations considered and the complexity of the road network. We consider every Census tract in the United States – upwards of 70,000 cells. Others have preceded us with finer granularity or national scale. For example, Li et al perform a county-level calculation for the continental United States. McGrail and Humphreys calculate access to primary care in Australia, using cells with just a few hundred people in rural areas. But the Australian road network is far less complex and the total number of origins is less than a third of ours.} But demand is still apt to permeate into or out of the catchment if supply is high or low. The current “state of the art” is the two-stage floating catchment area (2SFCA) method.\footnote{We base this statement on the total number of individual origin locations considered and the complexity of the road network. We consider every Census tract in the United States – upwards of 70,000 cells. Others have preceded us with finer granularity or national scale. For example, Li et al perform a county-level calculation for the continental United States. McGrail and Humphreys calculate access to primary care in Australia, using cells with just a few hundred people in rural areas. But the Australian road network is far less complex and the total number of origins is less than a third of ours.} With 2SFCAs, physicians are modeled as serving patients within a time or distance-based buffer of their office locations, and patients amass the sum of fractional assignments around them. Call the physician supply \( s_\ell \) at office location \( \ell \) and patient demand \( d_r \) at residential location \( r \). The time buffer \( T_\ell \) around a residential or office location \( \ell \) is the set of locations reachable within time \( t_{\text{max}} \):

\[
T_\ell = \{ \ell' \mid t_{\ell,\ell'} < t_{\text{max}} \}.
\]
Then ratio of physicians supplied to patients is

\[ R_\ell = \frac{s_\ell}{\sum_{r \in T_\ell} d_r} \]

and the 2SFCA patient accessibility by residence is the sum over available offices,

\[ 2\text{SFCA}(r) = \sum_{\ell \in T_r} R_\ell. \]

This formulation was originally motivated as an application of gravity models. \[\text{[1]}\] [In that vein, it has also been motivated as a logical successor of the Huff model.]

An approach similar to the 2SFCA method is to use two-dimensional kernel density estimation to model provisioned physician supply, \[\text{[2]}\] This approach has the notable relative disadvantage of supplying physicians isotropically rather than according to actual patient demand. This means that empty or underpopulated areas – parks, rivers, or low-density neighborhoods – will be modelled as over-provisioned with supply that will go “wasted.”

But the 2SFCA method is, in its base form, too simplistic. \[\text{[3, 4]}\] It is insensitive to the distances that patients must travel to receive care. 2SFCA implies constant use and constant utility for any time within the bounds \(t_{\ell r} \leq t_0\), and no use for \(t_{\ell r} > t_0\). A doctor at the limit of a patients’ travel band is used as much as one directly across the street. A result of this is that the population-averaged accessibility for the entire population is equal to the global PPR.\(^2\) This might at first appear rational, but it implies that the average spatial accessibility of care is independent of the spatial distribution of providers.

### 2.1.3. Current State of the Art: Enhanced Floating Catchment Methods

The Enhanced 2SFCA method (E2SFCA) begins from the basic insight that patients are more likely to use physician offices close to their homes than those that are further away. \[\text{[5]}\] In E2SFCA, locations are weighted by a decreasing function of travel time, \(W(t_{\ell r})\). This function is commonly specified as bands of time with, for instance, \(W_0\) for \(t_{\ell r} \leq 10\) minutes, \(W_1\) for \(10 < t_{\ell r} \leq 20\) minutes, and so forth. The values \(W_i\) may be derived from a Gaussian distribution. Aside these weights, E2SFCA mirrors 2SFCA exactly. One first calculates patient demand at physician locations \(\ell\) as the sum of demand from nearby residential locations \(r\), and then aggregates physician supply for each patient. Reusing our notation for physician supply and patient demand, the supply to demand ratio per office and the patient accessibility are:

\[ R_\ell = \frac{s_\ell}{\sum_{r} d_r W(t_{\ell r})} \quad \text{and} \quad \text{E2SFCA}(r) = \sum_{\ell} R_\ell W(t_{\ell r}) \]

The 2SFCA and E2SFCA methods share the weakness that the demand by patients on physicians does not diminish as a function of the number of available sites. Patient demand on each single doctor is independent of the available alternatives. While there is evidence \[\text{[6]}\] that patients visit the doctor more frequently in oversupplied locations, it is not credible to suppose that patients in a one-doctor town will begin visiting the doctor twice as often if a new physician moves in. A recent, compelling extension of E2SFCA adds an additional stage to address this. In the three-stage floating catchment area, patients distribute their own demand according to a distance-based allocation function, \(G_{r\ell}\). This function is defined as the ratio of the appeal of the location \(\ell\), \(W(t_{\ell r})\), with respect to the alternatives:

\[ G_{r\ell} = \frac{W(t_{\ell r})}{\sum_{r'} W(t_{r'\ell})}. \]

In the original 3SFCA method, \(W(\cdot) = W(\cdot)\); the same Gaussian function is used for both weights. What is curious about this mathematical structure is that it accounts for distances but not for the supply at each location \(\ell\). Splitting a single physicians’ group into two colocated practices would

\(^2\)Strictly speaking, this is true only if all doctors see at least one patient.
change this selection weight. At any rate, the accessibility then follows through two sums, exactly as before:

\[ R_\ell = s_\ell \sum_r d_{r\ell} W(t_{r\ell}) \quad \text{and} \quad 3\text{SFCA}(r) = \sum_\ell G_{r\ell} R_\ell W(t_{r\ell}). \]

It is worth noting that this modification represents different implicit assumptions about the elasticity of patient demand with respect to physician supply.

The interpretation of the E2SFCA and 3SFCA is identical to that of the nominal 2SFCA: it is the number of physicians serving each patient. However, past work has shown significant changes in modelled physician accessibility from changes in the specification of patients’ willingness to travel, \( W(\cdot) \). Since \( W(\cdot) \) is not in fact known, a number of analysts have advocated focusing not on the index of accessibility itself (E2SFCA or 3SFCA) but on its ratio with respect to the mean level of access – the spatial access ratio, or SPAR. They show that the SPAR is far more stable with respect to changes in \( W(\cdot) \). We continue this SPAR-based approach in what follows.

By modelling the decreasing utility of distant locations and incorporating a relative preference among locations, the E2SFCA and 3SFCA improve substantially on the 2SFCA. The models are quite general and extensible. The specification of the distance dependence \( W(\cdot) \) is quite flexible. Armed with travel time matrices on multiple modes, the procedures can be straightforwardly modified to allow separate populations to utilize separate modes. Patient demand per location can be modified to account for the unequal medical needs of different subpopulations, by replacing raw counts of patient with more-accurate models.

Still, some fundamental limitations remain. Neither E2SFCA nor 3SFCA incorporate patient responses to congested locations. Since this congestion is the very property identified as “poor access” at a global level, we contend that patients would choose to avoid it if possible. At the most mechanical level, oversubscribed physician practices will simply decline new patients. Stated slightly differently, floating catchment methods do not model any patient response to the initial allocation of demand nor any interplay between patient decisions. Our model, to which we now turn, is derived from this basic, competitive impulse.

2.2. The Rational Agent Access Model (RAAM). Our approach begins with the intuition that patients choose the physician practice that minimizes their combined access (travel time) and availability (doctor’s office congestion) costs. According to this rule, individuals shift their demand towards the “cheapest” point of care, accounting for others’ choices.

2.2.1. The basic model. We again denote the fixed supply of physicians at location \( \ell \) by \( s_\ell \), and the travel times by \( t_{r\ell} \). However, patient demand has an explicit destination as well as origin; \( d_{r\ell} \) represents the demand by residents of \( r \) at \( \ell \). The cost of care is defined as the sum of the congestion and travel costs. The “congestion cost” is the observed inverse PPR at \( \ell \) (patients per provider), accounting for demand by residents from all residential locations, and normalized by a factor \( \rho \): \( \sum_r d_{r\ell}/s_\ell \)/\( \rho \). The normalization \( \rho \) is fixed as the national inverse PPR, which was 1315 in 2010. The travel cost is simply the time \( t_{r\ell} \) from an agent’s residence \( r \) to \( \ell \) normalized by a parameter \( \tau \). This parameter sets the cost or disutility of travel relative to congestion. The total cost for a resident of \( r \) to receive care at \( \ell \) is thus

\[ \text{RAAM}(r, \ell) = \frac{\sum_r d_{r\ell}/s_\ell}{\rho} + \frac{t_{r\ell}}{\tau}. \]

Note that a common scaling of \( \rho \) and \( \tau \) simply scales all costs – the cost is homogeneous of degree one in \( \rho \) and \( \tau \). It is the relative values of \( \rho \) and \( \tau \) that affect relative access costs, so we are free to fix \( \rho \) and treat \( \tau \) is the sole free parameter. This makes \( \tau \) the single tunable parameter of the model; we show below that relative costs across locations are consistent over a broad range of \( \tau \). This account differs from 2SFCA and its derivatives by treating travel time as an explicit cost instead of a weight on care available from distant locations.

Just as important, this requires a model for \( d_{r\ell} \) – the locations at which patients actually seek care. The basic intuition of RAAM is that information about \( d_{r\ell} \) is already embedded in the cost function.
Assuming that patients seek care at the cheapest location, the choice of an agent at \( r \) can be expressed by the decision rule

\[
\text{argmin}_{\ell} \left[ \text{RAAM}(r, \ell) \right].
\]

In other words, patients choose the location with the lowest cost. Like most simple economic models, RAAM is predicated on patients having perfect knowledge of costs in the marketplace. Serban et al appeal to identical intuition in their decentralized optimization framework. As in their case, an optimization procedure is used to derive \( d_{r\ell} \).

This procedure treats home locations as “agents” aiming to equalize (and minimize) costs across locations. We begin with all patients assigned to the physician location closest to their home. We then cycle over residential locations, shifting demand from the most expensive used physician location to the cheapest available one. Patients, who are indivisible, are shifted by integer amounts. We calculate the patient shift necessary to equalize costs between the least and most expensive locations (see Appendix B). A shift that results in a full equalization may not be possible, since one cannot shift more patients from the expensive location than are presently there; at no time can \( d_{r\ell} < 0 \).

We also set an upper limit on the number of patients shifted per iteration so that the optimization proceeds gradually. The number shifted is thus the least of three values: (1) the cost-equalizing shift, (2) the current number of residents of \( r \) using the costly location, or (3) the configured upper limit. The algorithm cycles over residential locations and terminates when no reductions in cost are possible at any of them. The costs of a residential location \( r \) are the argument of this minimization when it is complete. At that point, that argument will be equal for all patronized offices of each residential location. In other words, for all minimum-cost locations \( \ell, d_{r\ell} > 0 \), RAAM takes on a single observed value,

\[
\text{RAAM}(r) = \frac{\sum_{r'} d_{r'\ell}/s_{r'}}{\rho} + \frac{t_{r\ell}}{\tau}.
\]

For any \( \ell \) with \( d_{r\ell} = 0 \), RAAM(\( r, \ell \)) \( \geq \) RAAM(\( r \)). Note that RAAM is anticorrelated with the FCA indices; RAAM represents costs whereas the previous methods assess physician availability.

It is worth unpacking this equation. Assuming homogeneous preferences across patients, all residents of each location should be indifferent among the actually-selected options (where \( d_{r\ell} > 0 \)). The net travel and congestion costs “paid” by patients of a single residential location must be equal at all physicians offices that they patronize. Patients may use different locations from their neighbors and they may have a different combination of congestion and travel costs. But because their costs are minimized, they cannot “pay more” for their care. For similar reasons, no non-utilized location could offer cheaper care; if it did, it would be utilized. By contrast, the patients of a single physician do not all face the same costs. They experience the same congestion at the point of care, but may travel different distances to reach it.

There are two important differences in the behavioral assumptions implicit in RAAM as compared with two-stage floating catchment methods. First is the elasticity of patient demand in response to supply and second is the relationship of patient choice and utility. In the floating catchment approach, each additional doctor results in higher accessibility for a patient. Patients in over-supplied areas receive more care (they aggregate a larger amount of fractional physicians). When supply increases, local patients consume more. This is as expected, but there is no saturation; the elasticity of demand is 1, which we think unlikely. By contrast, RAAM assumes that each patient consumes one unit of physician demand – it over-corrects this problem, with an elasticity of demand of 0. Outsiders are induced towards excess supply but local patients do not respond by consuming more. Note that the 3SFCA method, with its distance preference weights, is in this respect more similar to RAAM than to the (E)2SFCA.

In its simpler form, the 2SFCA method can be understood as a special case of the gravity or Huff model. This model treats usage as proportional to the inverse square of the distances from consumers’ homes. Importantly, usage is not concentrated at the closest possible location, and it is independent of other consumers. \cite{advocates} advocates this approach, arguing specifically that patient utility follows the
gravity model, and that utility and usage are proportional. This argument diverges dramatically from most economic models, where agents maximize utility or minimize costs. RAAM suggests that if one location is half as costly for a patient as another, it is strictly preferred – not utilized half as much. Distant locations are used only if the nearer ones are more costly; the distance dependence responds dynamically to the distribution of supply and demand. The downside of this perspective is that individuals do have real, independent reasons to leave the home – work, for instance – and these may make more-distant locations more attractive. A gravity or 2SFCA model implicitly incorporates this feature, while RAAM does not. We return to this issue and reincorporate it into the model, below.

In subsequent sections, we present a practical comparison of results from RAAM and floating catchment methods. But why pursue a new method at all? RAAM’s fundamental, conceptual advantages are two-fold. Foremost is the incorporation of the competition and interplay between patients into the heart of the method. Floating catchment methods do not incorporate a dynamic patient response to congested locations. These locations may be less valuable to patients (since they see higher demand, the supply to demand ratio is lower), but they are used nonetheless. RAAM naturally accounts for movement and what Li et al call “cascading” effects. At the edge of a region with high supply, patient location decisions will tend in that direction; this will alleviate local demand, gracing patients from further out with less congested care. In the language of the 2SFCA, a patient’s contribution to the supply to demand ratio of a location $R_i$ changes as a function of his or her outside options. Demand in RAAM is anisotropic: it depends on both distances and the availability of supply.

If physicians are added at a scarcely-accessible distance from a neighborhood, (E)2SFCA and 3SFCA both add these incremental options and improve residents’ accessibility even if the resources are not actually used. According to RAAM, these far-flung physicians reduce residents’ access costs only if the residents themselves or other patients of residents’ current doctors are induced to change locations. A well-supplied city will not place any demand on the over-booked country doctors of its hinter lands, and it is affected by those doctors only to the degree that they stanch the flow of rural patients towards the city. This description of patient decisions offers a different spatial logic for thinking about access. The alternate, implicit assumptions concerning patient choice and demand elasticity make RAAM a compelling counterpoint to FCA methods.

The second advantage is the separability of transit and congestion costs. Though individual patients at a location see different transit and congestion costs, their averages help illuminate the difference between high travel times and low provisionment of care. Large distances follow axiomatically from low density, but poor provider availability does not. To evaluate the drivers of poor accessibility in rural regions, one must be able to distinguish them.

Beyond these structural strengths, RAAM allows easy and deep extensibility. Further, its incidental outputs – the specific locations at which patients seek care – represent rich predictions that can be tested against (admittedly, limited) data. We next turn to these extensions and tests.

2.2.2. Modifying the model. As for floating catchment methods, many modifications of RAAM can be implemented by manipulating inputs or refactoring patient populations. Changes in patient response to travel can be applied as functions on raw travel times. A more accurate accounting of patient needs can be incorporated simply by replacing the raw population count with a demographic model for patient demand. We discuss below how to refactor patient populations to model multiple travel modes. With appropriate data, one could similarly construct networks of insurance or language comprehension.

Other modifications require deeper changes in the structure of the RAAM optimization. As an example, we modify RAAM to allow patients to “seek care” from either their home or work. Along similar lines, one might incorporate a demand response to costs. Seen as a whole, this section presents strategies for evaluating the systematic uncertainties on our findings.

Variable disutility of travel. A small literature explores patient responses to travel in their primary care utilization. These responses are captured in an array of weight functions in the FCA methods, and they are trivial to incorporate in RAAM. In each case, RAAM’s time cost ratio is replaced with
a function $f(t_{rt})$. The modified access cost becomes

$$\text{RAAM}'(r, \ell) = \frac{\sum r \cdot d_{rt}/s_\ell}{\rho} + f(t_{rt}).$$

As above, the (in)accessibility is the post-minimization cost of any used (minimum-cost) location $\ell$. RAAM'$(r, \ell) = \text{RAAM}'(r)$.

We suggest several alternatives for $f(\cdot)$, and present results in subsequent sections. The first is the square of the normalized cost, $f_1(t_{rt}) = (t_{rt}/\tau)^2$. The second takes the transform, $f_2(t_{rt}) = \log_2 (t_{rt}/\tau + 1)$. Both functions run between 0 and 1 at $(t_{rt}/\tau) = 0$ and 1, but the squared cost models each additional minute in the car as more burdensome than the one before it, while the logarithm treats it as less so. Along similar lines, research has suggested that rural populations have higher willingness to travel for medical care. [?] To model this, we define $\tau_r$ as a linear function of the fraction of the county population that is rural, $f_R$, running from 45 to 75 minutes: $\tau_r = 45 + 30 \times f_R(r)$ minutes. This treatment can also be considered as a systematic uncertainty if our travel costs are biased low in cities: it makes urban travel “more costly.”

**Multiple modes of transportation.** [cite papers provided by reviewers, new paper by Lin, Wan, et al] [People who take public transport have worse access.] This concern has been addressed in several recent floating catchment analyses. Mechanically, one simply replaces patients’ place of residence with the joint identifier of location and travel mode. Patient demand for each travel mode corresponds to an analytically distinct though spatially coincident residential population, with different travel times to doctor offices. All that changes are the number of locations and their populations, and the corresponding dimensions for the travel time matrix. The same approach applies for RAAM. The challenge for this analysis is to calculate travel times on public transportation at scale. That calculation is described in Section 3.3.

**Multiple origin locations.** [activity space literature] We are aware of a single paper that applies this commuting logic to a “Commute-Based 2SFCA (CB2SFCA),” in the context of childcare centers. Their strategy in short, was to treat the catchment area over which demand is aggregated as not a circle but an ellipse, so that the detour to visit a location on the way from home to work was less than some threshold $t_d$: $t_{hd} + t_{\ell w} < t_{hw} + t_d$. Implementing this requires data on the work locations of every resident.

We take a slightly different approach for RAAM. Unlike visits to childcare centers, doctors appointments are not quotidian activities and they are not confined to rush hours. We structure the agent’s choice not as an ellipse but as a two-part decision: first, whether to seek care from work or from home, and second, the location to patronize. Patients who seek care from work receive care for the same cost as patients who live there, and their demand is reallocated to originate from that location. The agent’s choice becomes:

$$\argmin_{r \in \{h,w,\ell\}} \left[ \delta(r = h) \times \left( \sum d_{ot}/s_\ell + t_{rt}/\tau \right) + \delta(r = w) \times \text{RAAM}(w) \right],$$

where $\delta(r = x)$ is an indicator function, valued 1 if $r = x$ and 0 otherwise. The summation is over home and work origin locations, so that demand stems not only from residents but also workers. This changes somewhat the mechanisms of RAAM, because neighbors with different workplaces face different costs. Workers who endure long commutes may find healthcare less costly near their work than at their home. Still, these long commutes and the consequent decisions to visit the doctor near work alleviate demand near home. The cascading effects are still in force.

Implementing this dual-origin approach requires, as for the CB2SFCA, an accounting of the home and workplace locations for all workers in the United States. Our source is the LEHD Origin-Destination Employment Statistics (LODES), discussed in the next Section. Using these data, we create a network of “tunnels” between residential locations whose capacity is determined as the actual number of workers. Demand shifts back and forth along these tunnels, as homes and workplaces offer
lower costs of care. (Non-workers do not have a workplace option.) The number of residents at \( r \) is \( N_r \) and the number opting to seek care at work is \( W_r \). The set of workplaces or “tunnel destinations” is \( \mathcal{W} \), and the number of users of tunneling from \( r \) to \( w \) is \( n_{rw} \). The average costs for residents of \( r \) can then be expressed,

\[
\text{Dual-Origin RAAM}(r) = \left[ (N_r - W_r) \times \text{RAAM}(r) + \sum_{w \in \mathcal{W}} n_{rw} \times \text{RAAM}(w) \right] / N_r.
\]

2.2.3. Validation from Data (PCSA). Models of human behavior stand and fall on their conformity with data. Work on floating catchments most frequently confront data at the stage of the specification of the patient distance response function, \( W(\cdot) \). Some work like [?, ?,?] to check which weights functions, If low local accessibility spurs patients to drive further as RAAM would suggest, a single distance response will not be appropriate. The output values of accessibility are seldom confronted with data.\(^3\) Data on actual accessibility may be somewhat subjective [ ] and is available to few researchers..

One way to confront data is to examine the location decisions of patients. Data on these locations is somewhat limited, but the Dartmouth Institute calculates a “preference index” that represents the fraction of care sought in the home region (Primary Care Service Area, discussed below). The “preference index” is embedded in RAAM’s demand matrix \( d_{rt} \). In the floating catchment approaches it is implicit in the physician locations from which patient “doctors per patient” is aggregated. Call the region of a location \( R_\ell \), and again denote the indicator function by \( \delta(\cdot) \). Then the preference fractions for the three models are:

\[
\mathcal{F}_{\text{E2SFCA}}(r) = \sum_\ell R_\ell W(t_{rt}) \delta(\mathcal{R}_r = \mathcal{R}_\ell) / \text{E2SFCA}(r),
\]

\[
\mathcal{F}_{\text{3SFCA}}(r) = \sum_\ell G_{rt} R_\ell W(t_{rt}) \delta(\mathcal{R}_r = \mathcal{R}_\ell) / \text{3SFCA}(r), \text{ and}
\]

\[
\mathcal{F}_{\text{RAAM}}(r) = \sum_\ell d_{rt} \delta(\mathcal{R}_r = \mathcal{R}_\ell) / \sum_\ell d_{rt}.
\]

A weighted “least squares” error can then be calculated for the model’s location predictions with respect to the PCSA data. Note that \( \mathcal{F}_{\text{E2SFCA}}(r) \) and \( \mathcal{F}_{\text{3SFCA}}(r) \) are ill defined if the allocated access is 0. In that case, we define care as outside of the home region.

For dual-origin RAAM, this calculation is complicated slightly, because it is not known if an agent who seek care from their workplace find it in their home region. We assume that workers who work in their home region and seek care from work have an equal likelihood to residents of the work location, of selecting a physician location within the home region. On the other hand, we treat patients who seek care from a workplace that is outside of their home region as seeking care outside of that region.

We present this consideration symbolically in an Appendix.

2.2.4. Computational implementation. The RAAM optimization algorithm is implemented in c++ with bindings to Python. The code is open-sourced and freely available for download. In what follows, we describe a Census tract level travel time matrix for the United States that includes all patient-provider pairs within a 100 km radius of each other – approximately 120 million pairs. (That matrix is also available for download.) Running the base model of RAAM with that matrix requires about 4 GB of memory. On a moderately powerful laptop, the optimization takes a few minutes.

3. Data and Technical Calculations

Modeling healthcare accessibility requires two inputs: locations of patients and providers, and a matrix of travel times between locations. This matrix has long been derived using expensive software or APIs. We demonstrate that these costs are avoidable by deriving matrices of unprecedented scale with open source tools deployed on inexpensive cloud computing.

\(^3\)Inverting 2SFCA to examine hospital congestion instead of physician availability, [ ] contrast their predictions with actual patient visits. [Li et al...]

Modifying our model and contextualizing its outputs also requires standard socioeconomic covariates, two measures of rurality, and a less commonly used Census data resource.


3.1.1. The Primary Care Service Area File. Data on primary care physicians is derived primarily from the Census tract-level Primary Care Service Area (PCSA) file for 2010, prepared for the Health Resource Services Administration by The Dartmouth Institute (TDI). We extract from this file the number of general practitioners per tract (\( TG_{DOC} \)), which itself is derived from Masterfile of the American Medical Association (AMA). The AMA Masterfile is a complete, administrative census of American physicians.

The PCSA also includes TDI’s estimate of localization of primary care – the so-called “preference index” for patients’ own PCSA (\( PF_{NDX} \)). TDI’s PCSAs are regions defined to optimally encapsulate primary care utilization behaviors, as already mentioned. \([?]\) These regions are derived using claims data from Medicare Part B beneficiaries. The preference index records the fraction of primary care received in-PCSA, by this population. It is worth emphasizing that these beneficiaries are a distinctive subpopulation that may be expected to have lower mobility and therefore a higher preference fraction than the broader population.

3.1.2. The US Census and the American Community Survey (ACS). Patient population counts are from the 2010 US Census at the tract level. For the multi-modal variant of RAAM described above, 5-year estimates for the American Community Survey are used to specify the population constrained to travel by public transportation. This is defined as the number of households without access to a car, weighted by the population size. ACS data are also used for the covariates of the regressions presented in Section 5. The last of these regressions considers physicians at their place of residence, who are also extracted from the Integrated Public Use Microsamples (IPUMS) of the 2010 ACS. \([\text{Physicians are required to be working full time, to have a graduate degree, and to be between 30 and 65 years old.}]\)

3.1.3. LEHD Origin-Destination Statistics (LODES). The dual-origin model previously described requires estimates of the work locations. These are drawn from the Census’ Longitudinal Employer-Household Dynamics (LEHD) Origin-Destination Employment Statistics (LODES). \([?]\) These data are drawn from state unemployment insurance records and cover approximately 95 percent of wage and salary jobs. For the year of this study (2010), Massachusetts was not contained in the file.

3.1.4. Measures of rurality. We use two county-level measures of rurality. The US Census classifies each tract in the nation as either urban or not. Aggregating this classification yields a county-level measure of the fraction of the population that is urban or rural. This is used for the definition of a rurality-dependent aversion to travel, \( \tau_r \). We also use the rurality classification of the National Center for Health Statistics (NCHS), to present the dependence of accessibility to care on rurality.

3.2. Network Data: OSM/GTFS. \([\text{Add short section on GTFS and OSM data sources}]\)

- OSM
- GTFS

3.3. Travel Times at Scale. To construct the origin-destination (OD) travel time matrix, we built a distributed data pipeline that utilizes free and open source software (FOSS), publicly available data sources, and cloud computing. The pipeline quickly, inexpensively, and accurately calculates the travel times between millions or even billions, of origin-destination pairs. It accounts for differences in travel speeds between urban and rural areas, and can incorporate multiple transportation modes. Combined, these capabilities significantly decrease the effort and cost required to create OD matrices, which were previously computationally limited to small geographic areas or available only to large, well-resourced companies. A matrix of this size and granularity would cost over half a million dollars through the

\(^4\)Verify that it’s not VT_NDX.
Google Maps API, and the license would not allow caching this result. A single term license of ArcGIS Network Analyst costs $600 and – on a single machine – would be incapable of performing the calculation. Our pipeline can perform the necessary computation for less than $20.

The pipeline computes a matrix of shortest-time travel times by routing from each origin to all destinations within a configurable range. For our national case study, we use population-weighted centroids of Census tracts as both origins and destinations. Each county is buffered to 100 km, and all origins within a county are routed to all destinations within the county and its buffer. Routing is performed by the PostGIS extension pgrouting using a national road network from OpenStreetMap (OSM) [?], cleaned and extracted using osmium, and loaded into Postgres with osm2pgrouting. At a national scale, this process is very slow; loading the entire North American OSM road network into PostGIS and then routing between all relevant points takes weeks on a single computer.

We achieve an acceptable calculation time by subsetting the routing into discrete jobs, which can be submitted to a custom Docker image deployed on a cloud computing service, such as Amazon Web Services (AWS). Each U.S. county is an independent job: for all tract centroids within a county, we calculate the driving times to all other tract centroids within the county and its 100 km buffer. By distributing the calculation across thousands of nodes, we reduce the time needed to calculate a national matrix to about 2 hours. The resulting matrix has full coverage: each individual origin will have travel times to all destinations within 100 km.

We do not include any node impedance (intersection delays), and we apply no penalty for crossing (sub-national) political boundaries. However, we do distinguish rural and urban driving speeds, which allows us to simplistically account for differences in traffic and speed norms. Speed limits are set for each OpenStreetMap road type, in urban and rural areas, considering average travel speeds. These settings yield trip times that are generally within 25 percent of the times returned by commercial mapping services – Google, Bing, and HERE – for the same trip. Because we do not model traffic, our results tend to underestimate the time required for short trips in large cities such as New York and Philadelphia (see Appendix A).

The pipeline also computes travel times for other modes of transportation. Using the open-source routing software OpenTripPlanner (OTP), we calculate the travel times between OD pairs using a combination of walking and public transportation (buses, subways, light rail, and heavy commuter trains). This method is applicable only where significant public transportation networks exist and where scheduling data is available (in a General Transit Feed Specification format). As such, we calculate public transit travel times only for the 40 largest U.S. transit systems, by ridership. As for driving, OTP calculations can be containerized and distributed across many compute nodes. With each city running in parallel, travel times on the 40 largest transit systems can be calculated in a few hours.

RAAM is free to use, but this essential input – travel times – often poses a significant computational burden. The sophistication of open source tools for network routing has increased dramatically over the past several years, and we hope our work and code will stimulate their use in this context. Nevertheless, since the cloud computing is a significant undertaking for a small project, we have taken pains to ensure that our output matrices are easily available and accessible to other researchers.

4. Access Results

The modelled accessibility of primary care is shown in Figure ???. Results are shown as a fractional deviation from the national mean, RAAM. In other words, it is the spatial access ratio (SPAR) minus one: RAAM(r)/RAAM – 1. The scale and granularity of this calculation, understood as the total number of patient and provider sites, is unprecedented in the existing literature. By analyzing the entire country, we have eliminated edge effects and enabled national comparisons.

This result sets the τ parameter to 60 minutes, which means that a 10 percent increase in the demand on a physician is commensurate with a 6 minute increase in travel time. This may seem like a steep price in time for a small increase in congestion at the point of care, but it is worth noting that after reaching full capacity, physicians’ supply is likely to be quite inelastic: prices for care are fixed.
per service, and there are finite hours in a day. More to the point, Figure ?? shows that the results are insensitive to \( \tau \) over a broad range of reasonable values. That plot shows the Spearman’s rank correlation coefficient between results derived with different values of \( \tau \), from 15 minutes to 3 hours. Results derived with \( \tau = 60 \) minutes are correlated \( \rho > 0.9 \), with results derived from \( \tau = 20 \) minutes to \( \tau = 3 \) hours. The choice of \( \tau \) affects the total (unit-less) cost by making long trips (even) less palatable, and affects regions’ costs relative to the national mean. But it has little effect on the accessibility rank.

The central aim of this project is to understand deficits in the accessibility of care, especially in rural America. Figure ?? shows the normalized cost distribution (z-scores) by rurality, as defined by the National Centers for Health Statistics (NCHS). There is a striking progression of costs from major metropolitan areas and their fringes, through medium and small metropolitan areas, to micropolitan and rural counties. As already suggested by Figure ??, this behavior is robust against different choices for \( \tau \).

Figure ?? presents the alternative models of Section 2.2.2, in the same format as Figure ???. It shows non-linear functional forms for the patient response to travel time, and multiple travel modes or origin locations. Each of these modifications affects the distribution of access costs without altering the broad conclusion of progressively higher spatial costs in rural areas. Perhaps the most notable difference is that the \( (t/\tau)^2 \) and dual-origin models both show a compression in the cost distributions of major metropolitan areas and their fringes. In essence, each of these sets the cost of short trips to 0. This is consistent with the deweighted temporal access costs seen for \( \tau = 90 \) minutes, in Figure ???. By contrast the model with rurality-dependent willingness to travel, \( \tau_r \), shows a compression between the most- and least-rural areas.

Car ownership is high outside of cities, and we calculate public transportation travel times only within large metro areas. The multi-modal model therefore has negligible impact outside major cities. To show these effects, Figure ??A-C zooms in to Cook County, Illinois, where Chicago is located. Travel times are longer on public transportation than driving (A). This means that failure to account for populations’ mobility options tends to underestimate costs (B). The brunt of these costs are borne by those populations without access to a car. In Chicago, the effect is therefore localized on the poorer South and West Sides (C).

This confirms what might have been intuited, namely, that populations limited to slower travel options have higher access costs for care. Disadvantaged populations face higher costs. However, these differences are small when compared to variation at the national level. Even for the affected populations, the increase is less than 20 percent of the average national costs (for \( \tau = 60 \) minutes), which in Chicago are far below the national average (see Figure ??A). Past work has documented the many non-spatial barriers that these populations face, which may be more significant. [? , ?]

We next contrast the results of RAAM with enhanced two- and three-stage floating catchment results (E2SFCA and 3SFCA). The distance decay in these models have tunable parameters that play the same role as \( \tau \) for RAAM. [Mention SPAR] We use Gaussian weights and three distance decay bands of 10, 20, and 30 minutes for 2SFCA. The original 3SFCA proposal used four bands of 10, 20, 30, and 60 minutes [ ]; it also requires a “preference” setting for closer locations, which we set to \( \beta = 640 \). At the national level, the weighted correlation between E2SFCA and 3SFCA is higher (0.70) than between them and RAAM (-0.64 and -0.65 respectively). Those correlations are shown in Figure ??A. Trends as a function of rurality are largely consistent. Figure ??B shows changes in quantile between RAAM and floating catchment methods. It is worth noting that the consistency of relative access in rural areas between RAAM and the FCA methods does vary with the chosen distance decay.

Finally, Figure ??A presents the difference between the measured preference fractions and those implied by the location allocations of RAAM \( \tau = 60 \) minutes, E2SFCA, and 3SFCA. Single-origin RAAM is shown to be biased high: too much care is in the home region. We postulate that this is due to the absence of any driver for patient care away from residence, as exists in reality. The home and work dual-origin model encodes this mechanism, and we therefore also show this model. This mode shows substantially lower error than the single-origin model. We again emphasize that evaluating the
dual origin model requires the assumption that if patients seek care from a work location outside of the region of residence, the care location is also outside of that region (PCSA). Figure ??B shows the sum of these squared “errors,” as a function of $\tau$. This shows that the RAAM error falls with increasing $\tau$; although slope evens out at $\tau \geq 60$ minutes, it does not reach a minimum. So how does RAAM fare? Our simple suggests that non-home activities should be taken into account, but that with this modification RAAM performs comparably with the FCA models.

5. The Determinants of Access

The findings of the preceding section illustrate the known shortages of primary care in rural America. [?, ?] Region by region, cities have lower access costs for care than their hinterlands. Our modified models show fairly consistent results for rural populations.

Notwithstanding this initial impression of high rural costs, there are enormous heterogeneities in access across rural areas. One can see by eye that Vermont has lower access costs than Utah or Texas. What drives this? We pursue an analysis at the level of the Public Use Microdata Areas (PUMAs). A PUMA is a density-dependent geography encompassing between 100 and 200 thousand people; the population scale is comparable to that of U.S. counties, but it is less variable. Controlling for the logarithm of the PUMA population density, Alaska has the lowest costs, but it is a special case: its road network is not connected and it has extraordinarily low density. It is followed by New England (Vermont, Maine, New Hampshire, and Massachusetts) and the northern plains states (Montana and the Dakotas). Meanwhile, costs are higher for inhabitants of the deep south (decreasing from most severe: Texas, Mississippi, Georgia, Louisiana, and Alabama) and in the lower mountain states (Utah, Nevada, and Arizona).

To separate the impact of density per se from that of other socioeconomic drivers, we present several regressions at the PUMA-level of the form,

$$\text{access} \sim \text{density} + \text{education} + \text{etc.}$$

We focus our exposition on the bachelor’s degree attainment of the adult population, which is the socioeconomic observable with the largest explanatory power. Table 1 presents three sets of regressions, contrasting each time the relative explanatory power of density and population education. The outcome variable changes from one set of regressions to the next: (1) the full RAAM cost ($\tau = 60$ minutes), (2) the average congestion cost – the inverse PPR at the point of care, and finally (3) the number of physicians per thousand residents. The last of these draws the physician counts from the IPUMS samples rather than the PCSA file. from the Integrated Public Use Microdata Series (IPUMS) of the 5-year files of the American Community Survey for 2010. [?] The aim of these specifications is to “peel back” the levels of commuting. The RAAM cost is the combined travel and congestion costs that patients experience. The average congestion costs allow for the fact that while long travel times may be an axiomatic consequence of rural life, poor physician availability is not. The congestion quantifies the physicians serving patients in each region, allowing for differences in patient travel across regions. Finally, the resident physicians eliminates travel by both patients and physicians; it is proportional to the propensity of a resident to be a physician.

Density itself explains the largest fraction of variance for the full-RAAM model ($R^2 = 0.37$). Adult educational attainment substantially improves the predicted accessibility: $R^2 = 0.28$ on its own and 0.49 in tandem with population density. Constraining the sample to the five hundred least-dense PUMAs, education explains 19% of the variance while the population density explains only 1%. Other socioeconomic covariates are less powerful. Turning to the congestion at the point of care, we find that though density remains strongly predictive of high costs its explanatory power is equal to that of bachelor’s degree attainment. Finally, to drive the point home, we identify physicians’ residential

\footnote{We tested higher-order polynomials of the non-logged population density, but the polynomial specification saturated at 6th order with $R^2 = 0.34$ of the variance whereas log population density singly explained $R^2 = 0.37$.}

\footnote{Note that unlike the combined costs, the two components of the RAAM cost are not constant at each residential location: some residents travel more for lower congestion while their neighbors travel less but have higher congestion.}
locations instead of offices. In this case, the $R^2$ for log density plummets to 0.07, while the fraction of adults with a bachelor’s degree leaps to $R^2 = 0.53$. Doctors live in educated areas; rurality is secondary.

Our interpretation of these results is that doctors live in and serve better-educated regions, which are often urban. They can commute away from these areas only to a limited degree. This is worth emphasizing: the focus on a “rural shortage” of primary care is potentially a red herring. It is true that rural areas are underserved. But there is enormous heterogeneity across rural areas and it appears that education is in some respects a better predictor of a shortage of doctors. [Cf McGrail on heterogeneity.] There is an important distinction between policies geared towards rural doctors and those geared towards underserved populations.

6. Discussion

Figure ?? presents the main result of our demonstration: a map of spatial access costs in the United States. The largest outliers are the region along the Mexican border in the Rio Grande Valley, the national parks of southern Utah, and the “last frontiers” of Alaska. These are places with high fixed travel costs: it is expensive to travel between towns, regardless of their primary care availability. Smaller but still noticeable outliers are islands without causeways to the mainland. Our travel time matrix does not take into account ferries. For disconnected components of the road network, the physician to population ratio is fixed by the local endowments.

RAAM tends to result in smoother access maps than and E2SFCA or 3SFCA. It is not unusual to find 3SFCA results with high- and low-access Census tracts immediately adjacent to each other. With RAAM, this “arbitrage” opportunity is not possible. The largest possible difference between the costs in two tracts is the travel time between them. The national access map in Figure ?? does have a few local outliers, but these tracts are all islands or have extremely inaccessible centroids, for instance, near International Falls, MN.

- E2SFCA + 3SFCA correlations.
- Variants generally agree
  - Additional applications – optimization, Cf appendix.
- Data-driven validation: single-source v. multi-origin shows much smaller errors (need more data!!!)
- Limitations

Because their data inputs are identical, RAAM shares some of the limitations of floating catchment methods. First, in calculating the travel times, we treat the origin in each Census tract as its population-weighted centroid, computed from the block-level, as has been done in a number of previous articles. This procedure usually yields reasonable locations, but as just noted, it results in a few unnecessarily remote centroids when considering the entire country. In addition to International Falls, the points in Grand Escalante National Park in Utah or West Canada Lake Wilderness in upstate New York are less accessible than the average residences in those tracts. The fraction of points affected is however small.

The second limitation is in the estimates of demand for and supply of physicians. We have treated each individual as one unit of demand. We do not account for variability in patient needs as a function of age or sex. Similarly, we treat each primary care physician as a single unit of demand and ignore care provided by nurses or residents. Both of these limitations yield to the simple technical solution of reweighting care demanded or provided; indeed, reweighting provider capacity is the strategy advocated by Rickets et al for calculating Federal HPSAs. [?] With the public transportation model we incorporated variation in travel costs. But we are blind to variability in patients’ ability or willingness to overcome spatial or non-spatial barriers to care. Except for the high rural willingness to travel variant ($\tau_U \neq \tau_R$), the model as presented assumes homogeneous travel preferences. The model ignores costs other than travel time and congestion. To the extent that these costs can be aligned with a site of care or a population, they could be incorporated into the model, as we have shown. But if high
costs of any type cause individuals not to seek care, they will reduce demand on and congestion of the system (at least at the primary care level). RAAM does not model this fall-out. This is related to the point made earlier: RAAM ignores both the intensive and extensive consumption response to prices.

- Education v. density
  - This depends on a calculation of unprecedented scale ⊗ scope.

[main area that needs more writing is discussion section]

7. Conclusions

- Fundamental construction is more convincing: competition
  - Standard and non-standard modifications are straightforward

This project has presented a Rational Agent Access Model (RAAM) for calculating spatial accessibility of primary care in the US. RAAM accounts for competition between locations and feedback between patients’ decisions, in a way that fundamentally different from floating catchment methods. RAAM is a versatile framework. It is trivial to alter patient response to distance, but RAAM also allows for multiple patient origin locations or travel modes – modifications that have only recently matured in floating catchment methods.

Using RAAM, we measure the accessibility of care across the United States and reproduce the well-known rural shortage. Turning to urban populations constrained to the public transportation network, we measure the difference in accessibility costs for populations with and without a car. These differences are small with respect to the variation at the national scale, between urban and rural areas. Similarly, accounting for commutes to work results in relatively small comparative effects, concentrated on the fringes of major metropolitan areas. Our baseline results are reasonably consistent with floating-catchment methods. Interpreting these results nationally, we offer suggestive evidence that the rural shortage may be driven more by population educational attainment than by population density.

Our work is substantially enhanced by an ambitious technical approach to calculating travel time matrices, using distributed resources. This affords our project unprecedented scope and granularity. We hope that this distributed approach will help reduce costs for other analysts.
Appendix A. Comparison of travel times with commercial suppliers.

Figure ?? compares driving times from our pipeline to commercial providers. We have chosen five major cities and five (random) rural counties. In each we plot a random collection of tract-to-tract trips. Our pipeline does not account for traffic, and our times are very low in Manhattan (New York County, New York) and somewhat low in Philadelphia. Results from HERE seem to be consistently high, particularly in cities. With travel time data across cities – from taxi trips, for example – one could fit city-specific average travel speeds by road-type.

Appendix B. Optimization Implementation: Mathematical Derivation

The optimization implementation cycles over residential locations \( r \), shifting patients from the maximum cost physician location used by residents, \( \ell \), to the minimum cost location available to them, \( \ell' \). The number of patients actually shifted is the least of three values: (a) the number needed to actually equalize costs, (b) the number originally at the maximum-cost location, and (c) a configurable maximum shift size.

We represent demand by residents of \( r \) as \( d_r = \sum_{\ell} d_{r\ell} \). Demand at physician locations can be decomposed into demand by self (\( r \)) and others (\( \{r\} \)): \( \sum_{\ell'} d_{r\ell'} = d_{r\ell} + d_{r\ell'} \). Residents can alter their own allocations \( d_{r\ell} \) but they cannot affect others’ location decisions; \( d_{r\ell} \) and \( d_{r\ell'} \) are fixed in each iteration.

The physician supply of a location \( s_\ell \) and the times necessary to reach it \( t_{r\ell} \) are permanently immutable.

How much demand \( d_{r\ell} \) should be placed at the current minimum-cost location? Set costs equal between it and the maximum cost location:

\[
\frac{(d_{r\ell} + d_{r\ell'})/s_\ell}{\rho} + t_{r\ell} = \frac{(d_{r\ell} + d_{r\ell'})/s_{\ell'}}{\rho} + t_{r\ell'}
\]

\[
\frac{d_{r\ell} + d_{r\ell'}}{s_\ell} = (t_{r\ell} - t_{r\ell'}) \frac{d_{r\ell} + d_{r\ell'}}{s_{\ell'}}.
\]

For each single iteration, we consider the two supply locations in isolation. Their sum is thus fixed and known: \( d_r = d_{r\ell} + d_{r\ell'} \). Using this to eliminate \( d_{r\ell'} \) yields

\[
\frac{d_{r\ell} + d_{r\ell'}}{s_\ell} = (t_{r\ell} - t_{r\ell'}) \frac{d_r - d_{r\ell} + d_{r\ell'}}{s_{\ell'}}.
\]

Now isolate \( d_{r\ell} \) on one side:

\[
d_{r\ell} = \left( \frac{s_{\ell'} - s_\ell}{s_\ell + s_{\ell'}} \right) \left[ (t_{r\ell'} - t_{r\ell}) \frac{d_r + d_{r\ell'} - d_{r\ell}}{s_{\ell'}} - \frac{d_{r\ell}}{s_\ell} \right].
\]

This represents the values of \( d_{r\ell} \) and implicitly \( d_{r\ell'} \) that equalize costs between the \( \ell \) and \( \ell' \). This represents just one of three possible caps on the number of patients to shift. As already noted, we may want the optimization to proceed gradually, and we cannot move more patients from a physician location than are currently using it. Therefore, it may not be desirable or possible to equalize costs based on this single shift.

Appendix C. Dual-Origin Preference Fraction

Dual-origin RAAM does not track the location at which every patient receives care: it tracks local care decisions from the residence. Care sought at work is re-allocated to that location, and the decisions of residents and workers are not distinguished.

The Primary Care Service Area (PCSA) Preference Fraction records the fraction of residents who seek care within their home PCSA region. For the dual origin model, we approximate this preference fraction as the weighted average of (a) the fraction who receive care in the local region, for those who seek care at home; (b) the fraction who receive care in the local region, for those who seek care from...
a workplace in the same region as the residence; and (c) 0, for those seek care from a workplace that is not in the home region.

To express this mathematically, call the PCSA (region) of \( i \) \( R_i \), the indicator for a shared PCSA for \( i \) and \( j \) \( 1(R_i = R_j) \) and the fraction of local and visiting patients at \( i \) who receive care in \( R_i \), \( f_i \). The number of residents is \( N_i \) and the number opting to seek care at work is \( W_i \). The set of workplaces or “tunnel destinations” is \( W \), and the number of users of tunneling from \( i \) to \( w \) is \( n_{iw} \). Then the in-PCSA fraction of the residential location can be written:

\[
\text{Fraction In-Region} = \frac{1}{N_i} \left( (N_i - W_i) \times f_i + \sum_{w \in W} 1(R_i = R_w) \times n_{iw} \times f_w \right).
\]

**Appendix D. The Optimal Placement of Physician Supply**

Beyond identifying physician shortages, accessibility measures can be used to optimize placement of additional supply. Where does one physician do the most good? It is not a priori obvious how to balance accessibility and scarcity. A physician in an undersupplied region could be so remote that she serves very few patients. To calculate the optimal placement, we first run the simulation to allocate patient demand across the system. We then add a single physician at a location, reoptimize demand, and remeasure costs. The marginal value of a physician in a location is equal to the total difference in costs that they cause over the entire system.

We perform this procedure for each of the the 2351 Public Use Microdata Areas (PUMAs) defined after the 2010 Census (as opposed to the 2071 available at that time). The computational cost of evaluating these 2351 alternative locations is substantial, but at a more-restricted geographic scope, one could consider the value of any single Census tract. In each PUMA, we select the single tract with the greatest number of practicing primary care physicians as the location for placing an additional physician. This proxies “where a physician would be likely to go,” but alternatives are obviously possible.

The result – the marginal value of a physician – is presented in Figure ???. This shows a basic consistency with the access costs of Figure ???: the marginal value of a physician is low in most cities and in the Northeast, Midwest, Northwest, and Bay Area. In underserved locations, the fractional change in supply from a single doctor is larger. But physicians’ marginal value is also fairly low in the Dakotas, Nebraska, and Wyoming. These are areas with very high travel costs, which an additional physician does not fix.
Table 1. Accessibility costs and doctor supply, regressed on population density and socioeconomic characteristics, at the PUMA level.

<table>
<thead>
<tr>
<th></th>
<th>Total RAAM Cost</th>
<th>Congestion Cost</th>
<th>Physicians / 1k Residents</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Intercept</td>
<td>2.06*</td>
<td>1.27*</td>
<td>2.02*</td>
</tr>
<tr>
<td></td>
<td>(0.03)</td>
<td>(0.01)</td>
<td>(0.03)</td>
</tr>
<tr>
<td>Log Population Density</td>
<td>-0.07*</td>
<td>-0.06*</td>
<td>-0.07*</td>
</tr>
<tr>
<td></td>
<td>(0.00)</td>
<td>(0.00)</td>
<td>(0.00)</td>
</tr>
<tr>
<td>Bachelor’s</td>
<td>-0.96*</td>
<td>-0.65*</td>
<td>-0.85*</td>
</tr>
<tr>
<td></td>
<td>(0.03)</td>
<td>(0.03)</td>
<td>(0.04)</td>
</tr>
<tr>
<td>Log Median HH Income Poverty</td>
<td>0.36*</td>
<td>0.36*</td>
<td>0.36*</td>
</tr>
<tr>
<td></td>
<td>(0.03)</td>
<td>(0.03)</td>
<td>(0.03)</td>
</tr>
<tr>
<td>Black</td>
<td>-0.05</td>
<td>-0.05</td>
<td>-0.05</td>
</tr>
<tr>
<td></td>
<td>(0.03)</td>
<td>(0.03)</td>
<td>(0.03)</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.37</td>
<td>0.28</td>
<td>0.49</td>
</tr>
<tr>
<td>$N$</td>
<td>2071</td>
<td>2071</td>
<td>2071</td>
</tr>
</tbody>
</table>

Standard errors in parentheses; * $p < 0.001$. 
Figure 1. Spatial access costs for primary care in the United States as a fractional deviation from national mean, for $\tau = 60$ minutes. Note that while the largest deviations do exceed 100%, the legend does not.

Figure 2. Rank correlation of access costs across Census tracts, with varying setting for the trade-off $\tau$ between travel time and congestion weight. For $\tau = 20$ minutes to 3 hours, the rank correlations to results from our baseline setting of $\tau = 60$ minutes are over 0.9.
Figure 3. Spatial access by NCHS rurality, and for varying definitions of \( \tau \). After demeaning and normalizing each distribution (taking the z-scores), the different values of \( \tau \) show markedly consistent trends in access against rurality. The boxes show the median values and interquartile ranges per rurality level, and the whiskers show then 10th and 90th percentiles.

Figure 4. Comparison of the impact of modified models on conclusions about the accessibility of care in urban and rural regions. Rurality is determined by county, according to the NCHS definitions.
Figure 5. Differences in calculated access costs, between driving-based or multimodal travel, or between driving and public transportation only, in Cook County, Illinois. Travel times are longer on public transportation than by car, leading to higher costs for populations using transit (A). This leads to an underestimate of costs when using driving time matrices alone (B). These costs are borne by the populations without access to a car. In Cook County, multimodal RAAM implies lower accessibility for the poor populations on the South and West Sides of Chicago (panel C, red areas).

Figure 6. Accessibility costs from the Rational Agent Access Model are plotted against the two- and three-stage floating catchment methods, showing strong negative correlations (−0.65 and −0.64 respectively).
Figure 7. The difference in quantiles for accessibility costs versus (negative) two- and three-stage floating catchment accessibility, for Census tracts. The box shows the interquartile range and the whiskers show the 10th and 90th percentile, of the distribution of the cost changes, for each of the six rurality groups defined by the National Center for Health Statistics (NCHS). For both FCA methods we use three time bands of 20, 40, and 60 minutes, and for the 3SFCA results we set $\beta = 640$. [? , ?].

Figure 8. Difference between use modeled in Primary Care Service Areas, and the “Preference Fractions” (individual-weighted use) measured by the Dartmouth Institute through Medicare claims (panel A). The difference is significantly reduced by the use of a dual-origin model but the bias is not eliminated. The population-weighted sums of squared errors (panel B) do not reach minima with respect to $\tau$: they remain biased high.

(A) Population

(B) Weighted Error

\[ \tau = 60 \text{ min.} \]
Figure 9. Driving travel times from our distributed pipeline are shown against those from commercial providers, for five major cities and five (random) rural counties.
Figure 10. The marginal value of a single physician varies dramatically across the United States. The map shows the impact of single additional physician in each PUMA, on the average costs to all patients in the United States, scaled up by $10^5$. 