A RATIONAL AGENT MODEL FOR THE SPATIAL ACCESSIBILITY OF PRIMARY HEALTH 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 offers several advantages with respect to existing access frameworks. It incorporates feedback between patient decisions and endogenizes 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 account for 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. We compare the results to those from existing primary care access models.

1. Introduction

Primary care is a critical component of effective health systems. [1] Higher quality primary care is associated with better health outcomes, better patient experiences, and lower health inequity, at lower costs. [2–7] Both in the United States and internationally, rural areas face shortages of primary care professionals. [8, 9] Where are these shortages most acute, and how do regions compare?

Measures of patient "access" to primary care may be organized conceptually according to two dichotomies [10]: potential as opposed to realized access, and spatial determinants of care delivery as contrasted with aspatial ones. Spatial access refers to the geographic distribution of physicians and patients. It encompasses both accessibility and availability: travel costs paid by patients as well as the (numerical) adequacy of the physician supply at the point of care. This paper focuses on the measurement of potential spatial accessibility. It ignores aspatial considerations like affordability, acceptability, or appropriateness. [11–14] However, since measures of potential access should correspond to the actual possibility of *use* [13], we do briefly contrast our measures of potential access with data on realized use. The paper presents two methodological contributions, the first for modeling and the second technical.

Our modeling contribution is an intuitive and efficient framework for access measurement, that offers several advantages with respect to state of the art two-stage floating catchment area (2SCFA) method and its enhancements. [15–19] Those methods first assess the demand by patients on physicians and then allocate the fractional supply of physicians available to each patient. This setup does not allow for interactions between patients: avoidance of congested locations or preference for available ones. These preferences are the bedrock of our approach. We model patients as seeking care at the physician office with the lowest combined cost of travel time plus congestion at the point of care. An efficient optimization iterates over patient locations, allocating demand to the cheapest physicians. This endogenizes the trade-off between travel times and office congestion and introduces feedback between patients' decisions. The algorithm terminates when all patients at each residential location experience the same costs – when no improvements are available – revealing the "cost of care" for patients. This functional principle and computational approach follow those developed by Serban, et al. [20, 21] This realistic, competitive response to congestion has not been available in floating catchment methods.

We extend this principal into a framework, which we call the Rational Agent Access Model (RAAM). RAAM incorporates many of the conceptual refinements common in floating catchment methods. We

E-mail address: jsaxon@uchicago.edu.

Date: June 9, 2019.

show how changing willingness to travel affects results. Several recent papers explore the impact of multiple travel modes [22–25], 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, new to the medical access literature, although existing work does consider home and work poles in a floating catchment approach to childcare sites. [26] Taken together, these modifications offer a raft of realistic systematics. We present strategies for contrasting modeled results with data.

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.[1]

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 extensions of our approach, which we evaluate in subsequent Sections.

2.1.1. *The Provider to Patient Ratio.* The simplest measure of physician accessibility is the provider to patient ratio (PPR): the average number of (fractional) physicians available to a patient. 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. [28] But 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 behaviors. TDI calls these regions Primary Care Service Areas (PCSAs). [29] However, despite the aim of better encapsulation of care, the PCSAs remain permeable. TDI therefore also measures "re-allocated" patient demand, using claims data to identify patients' residences and unwind the permeability issue by assigning care where it is consumed. 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 catchment area by a time or distance buffer, and scan or "float" this buffer across the study region. One then evaluates the PPR within that "floating" catchment area (FCA). [16,30] But demand

[1]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. [20] McGrail and Humphreys calculate access to primary care in Australia, using cells with just a few hundred people in rural areas (US Census tracts have a few thousand). [27] But the Australian road network is far less complex and the total number of origins is less than a third of ours.

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. [15,31] 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_ℓ at office location ℓ and patient demand d_r at residential location r. The set of locations reachable from residential or office location ℓ within time $t_{\max}$ is its time buffer $\mathcal{T}_\ell$,

$$\mathcal{T}_\ell \equiv \left\{ \ell' \,\middle|\, t_{\ell,\ell'} < t_{\max} \right\}.$$

The ratio of physicians supplied to patients at ℓ is then

$$R_\ell \equiv s_\ell \Big/ \sum_{r \in \mathcal{T}_\ell} d_r$$

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

$$\text{2SFCA}(r) \equiv \sum_{\ell \in \mathcal{T}_r} R_\ell.$$

This was originally motivated as a derivative of gravity models [32], and has also been framed as an extension of the Huff model of trade areas. [33,34]

An approach similar to the 2SFCA method is to use two-dimensional kernel density estimation to model provisioned physician supply. [35] 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 provisioned with supply that may go "wasted."

In this unadorned form however, the 2SFCA method is too simplistic. [17,36] It is insensitive to the distances that patients must travel to receive care. 2SFCA implies constant use and constant utility for any location within a travel time $t_{r\ell} \leqslant t_{\max}$ and no use for $t_{r\ell} > t_{\max}$. A doctor located around the corner is used no more than one at the limit of a patients' travel band. 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 for the entire study area, the average 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. [18] In E2SFCA, locations are weighted by a decreasing function of travel time, $W(t_{r\ell})$. This function is commonly specified stepwise in bands of time with, for instance, W_0 for $t_{r\ell} \leqslant 10$ minutes, W_1 for $10 < t_{r\ell} \leqslant 20$ minutes, and so forth. The values W_i may be derived from a Gaussian distribution, power law, or exponential decay. Aside from these weights, E2SFCA mirrors 2SFCA exactly. One first calculates patient demand at physician locations ℓ 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

$$R_\ell \equiv s_\ell \Big/ \sum_r d_r\, W(t_{r\ell}) \qquad \text{and} \qquad \text{E2SFCA}(r) \equiv \sum_\ell R_\ell\, W(t_{r\ell}).$$

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 [37] 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 method (3SFCA), patients distribute their own demand according to a distance-based allocation

[2]Strictly speaking, this is true only if all doctors see at least one patient.

function, $G_{r\ell}$. [19] This function is defined as the ratio of the appeal of the location ℓ, $\mathcal{W}(t_{r\ell})$, with respect to the alternatives:

$$G_{r\ell} \equiv \mathcal{W}(t_{r\ell}) \Big/ \sum_{\ell'} \mathcal{W}(t_{r\ell'}).$$

In the original 3SFCA method, $\mathcal{W}(\cdot) = W(\cdot)$; 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 ℓ. Splitting a single physicians' group into two colocated practices would change this selection weight. At any rate, the accessibility then follows through two sums, exactly as before:

$$R_\ell \equiv s_\ell \Big/ \sum_r d_r \, G_{r\ell} \, W(t_{r\ell}) \qquad \text{and} \qquad \text{3SFCA}(r) \equiv \sum_\ell G_{r\ell} \, R_\ell \, W(t_{r\ell}).$$

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

The interpretation of the E2SFCA and 3SFCA is identical to that of the nominal 2SFCAs: it is the number of physicians serving each patient. However, past work has shown that the values of modelled accessibility are sensitive to 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, which they call the spatial access ratio, or SPAR. [38, 39] They show that the SPAR is far more stable with respect to changes in $W(\cdot)$. We adopt this relative-access approach in what follows, and proceed a step further, focusing on normalized distributions and ranks instead of a raw metric.

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 distance dependence $W(\cdot)$ allows trivial re-specification. Armed with travel time matrices on multiple modes, the procedures can be straightforwardly modified to allow separate populations to utilize separate modes. [22–25] Patient demand per location can be modified to account for the unequal medical needs of different subpopulations, by replacing raw counts of patients with more-accurate models. [27, 28]

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.[3] 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. We now turn to our agent-based model, which is derived from this fundamental, competitive impulse.

2.2. **The Rational Agent Access Model (RAAM).** We begin with intuition similar to Serban et al, that patients choose the physician practice that minimizes their combined access (travel time) and availability (doctor's office congestion) costs. [20, 21] According to this rule, individuals shift their demand towards the "cheapest" point of care, accounting for others' choices. 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 not avoided. Accounting for user choice allows for "cascading effects," where some patients' choices will relieve congestion for others. [20] 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

[3]Congestion may indicate higher-quality or more efficient service or amenities like foreign language comprehensions, but these go beyond the scope of this paper. In itself the congestion is a pure "bad."

congested care. In the language of the 2SFCA, a patient's contribution to the supply to demand ratio of a location R_ℓ changes as a function of his or her outside options. Demand in RAAM 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. On its own, the alternative conceptualization of patient choice and its implicit assumptions about demand elasticity to supply (discussed after the model), 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 easily allows deep and realistic extensions. This is important for assessing the robustness of results and building intuition for their weaknesses. Accordingly, where past work on floating catchment approaches have typically presented one modification of the method at a time, we instead present an array of realistic variations. The incidental outputs of RAAM – the specific locations at which patients seek care – represent rich predictions that can be tested against (unfortunately limited) data. We present the model mathematically before turning to these extensions and tests.

2.2.1. *The basic model.* As above, we denote the fixed supply of physicians at location ℓ by s_ℓ, and 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 ℓ. The cost of care is defined as the sum of the congestion and travel costs. The "congestion cost" is the observed inverse PPR at ℓ (patients per provider), accounting for demand by residents from all residential locations, and normalized by a factor ρ: $\left(\sum_{r'} d_{r'\ell}/s_\ell\right)/\rho$. The normalization ρ is fixed as the national inverse primary-care PPR, which was 1315 in 2010. The travel cost is simply the time $t_{r\ell}$ from an agent's residence r to ℓ normalized by a parameter τ. This parameter sets the cost or disutility of travel relative to congestion. The total cost for a resident of r to receive care at ℓ is thus

$$\mathrm{RAAM}(r,\ell) \equiv \frac{\sum_{r'} d_{r'\ell}/s_\ell}{\rho} + \frac{t_{r\ell}}{\tau}.$$

Note that a common scaling of ρ and τ simply scales all costs – the cost is homogeneous of degree one in $1/\rho$ and $1/\tau$. It is the relative values of ρ and τ that affect relative access costs. We are therefore free to fix ρ and treat τ as the sole free parameter of the model. We show below that relative costs across locations are consistent over a broad range of τ. 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, the travel time matrix, and the geographic distribution of supply and demand. Assuming that patients seek care at the cheapest location, the choice of an agent at r can be expressed by the decision rule

$$\operatorname*{argmin}_{\ell} \left[\mathrm{RAAM}(r,\ell)\right].$$

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. [20, 21] 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.[4] 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, the costs to residents of r will be equal (and minimal) across patronized physicians' offices $\underline{\ell}$ (with $d_{r\underline{\ell}} > 0$). The total RAAM cost is thus a single-argument function of the place of residence,

$$\mathrm{RAAM}(r) = \frac{\sum_{r'} d_{r'\underline{\ell}}/s_{\underline{\ell}}}{\rho} + \frac{t_{r\underline{\ell}}}{\tau}.$$

For any ℓ with $d_{r\ell} = 0$, $\mathrm{RAAM}(r, \ell) \geqslant \mathrm{RAAM}(r)$. 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. On the other hand, 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 effect is observed in practice [37], but there is no saturation in the model; the elasticity of demand is 1, which we think unlikely. RAAM instead 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 a Huff model with a gravity kernel. In this case, usage is not concentrated at the closest possible location, and it is independent of other consumers. Elaborating this approach, Wang and colleagues express patient utility as proportional with the gravity kernel, which is then proportional to usage. [34, 40] This approach differs markedly from standard economic assumptions, where agents maximize utility or minimize costs. RAAM suggests that if one location is costs twice as much for a patient as another, it is completely avoided – 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.

[4] Alternative initialization strategies yield consistent results.

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.

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: an elasticity of demand to supply. Seen as a whole, this Subsection presents strategies for assessing 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 $f(t_{r\ell})$. The modified access cost becomes

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

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

We suggest several alternatives for $f(\cdot)$, and present results in subsequent Sections. The first is the square of the normalized cost, $f_1(t_{r\ell}) = (t_{r\ell}/\tau)^2$. The second takes the transform, $f_2(t_{r\ell}) = \log_2(t_{r\ell}/\tau + 1)$. Both functions run between 0 and 1 at $(t_{r\ell}/\tau) = 0$ 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. [41] To model this, we define τ_r as a linear function of the fraction of the county population that is rural, f_R, running from 45 to 75 minutes: $\tau_r = 45 + 30 \times \text{f}_R(r)$ 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. Travel times vary significantly according to the transit mode. Especially in cities, poorer populations do not universally have access to a car. If they are assumed to use a "driving" network their travel costs will tend to be understated. This concern has been addressed in several recent floating catchment analyses. [22–25] To model this, one simply replaces patients' place of residence with the joint identifier of location and travel mode. Patient demand through 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 (analytical) 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 3.2.2.

Multiple origin locations. A well-developed geographic literature on "activity spaces" characterizes the poles of and temporal constraints on individuals' daily routines (see [42,43] for early papers or [44] for a modern review). We are aware of a single paper that applies this logic to floating catchment accessibility, a "Commute-Based 2SFCA" (CB2SFCA) measure for childcare centers. [26] The strategy of that paper was to model 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: $t_{h\ell} + t_{\ell w} < t_{hw} + t_d$. Implementing this requires data on the work locations of every resident.

We take a slightly different approach for RAAM. Unlike visits to childcare centers, appointments with a doctor are not quotidian activities and they are 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 home h or from work w, 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:

$$\underset{r \in \{h,w\},\ell}{\operatorname{argmin}} \left[\delta(r = h) \times \left(\frac{\sum_o d_{o\ell}/s_\ell}{\rho} + \frac{t_{r\ell}}{\tau} \right) + \delta(r = w) \times \text{RAAM}(w) \right],$$

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" for r is $\mathcal{W}_r$, 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,

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

2.2.3. *Comparing Modelled and Realized Use of Locations.* Though data have informed the specification of access models – the sizes of catchments [41] or the functional form of patient-physician interactions [40] – validation of outputs has been limited (though see [34]). There are two reasons for this. First, the output of a model of accessibility – the constructed concept of "potential access" – is not directly measurable. Health researchers have long argued, however, that potential access should bear a direct relationship with actual utilization of health services – that the proof of "access" is use. [12, 45] The second problem, then, is that data on realized access are themselves often sparse and inaccessible.

One way to bring data to the model is through its auxiliary outputs – the locations at which patients receive care. Patient office visits are recorded in Medicare Part B claims. Although individual-level data are not public, the Dartmouth Institute calculates and releases a "preference index" that represents the fraction of care that patients seek in their home region (Primary Care Service Area, discussed below). The *modelled* preference index of each region is embedded in RAAM's demand matrix $d_{r\ell}$. In the floating catchment approaches it is implicit in the physician locations from which patient "doctors per patient" are aggregated. Call the region of a location $\mathcal{R}_\ell$, and again denote the

indicator function by $\delta(\cdot)$. Then the preference fractions for the three models are:

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

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

$$\mathcal{F}_{\text{RAAM}}(r) = \sum_{\ell} d_{r\ell} \, \delta(\mathcal{R}_r = \mathcal{R}_\ell) \Big/ \sum_{\ell} d_{r\ell}.$$

The error in the home-region is the difference with respect to the empirical values. Note that the modelled fractions $\mathcal{F}_{\text{E2SFCA}}(r)$ 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.

Results with these data must be interpreted gingerly. The information available in the preference fraction is limited: it is either home or not. A model does not need to derive demand from the *right* locations in order to achieve the correct fraction, it simply needs to use and avoid the home location the correct amount.

For dual-origin RAAM, this calculation is complicated slightly, because we do not record if agents who seek care from their workplaces find that care 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 caveat 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 single core on a moderately powerful laptop, the optimization takes a few minutes. In other words, RAAM runs comfortably on networks of unprecedented scale. The processing time and hardware requirements are comparable with existing methods.

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.

Modifications of our model require Census measures of car ownership, work locations, and rurality. To contextualize our findings, we also employ a second measure of rurality.

3.1. **Population Data.**

3.1.1. *The Primary Care Service Area File.* Data on primary care physicians are 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 is itself derived from the 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. [29] These regions are derived using claims data from Medicare Part B beneficiaries. The preference index records the fraction of beneficiaries who

prefer (patronize) providers in this PCSA. 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 size of the household.

3.1.3. *LEHD Origin-Destination Statistics (LODES).* The dual-origin model previously described requires data on residents' work locations. These are drawn from the Census' Longitudinal Employer-Household Dynamics (LEHD) Origin-Destination Employment Statistics (LODES). [46] 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, τ_r (see Section 2.2.2). 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. **Transportation Networks and Travel Times.**

3.2.1. *Network Datasets.* Calculating travel times requires two types of datasets. First is the OpenStreetMap driving network for the United States. [47] The second are public transportation schedules and networks, in the General Transit Feed Specification format.

3.2.2. *Travel Time Calculations.* 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. For our national case study, we use population-weighted centroids of Census tracts as both origins and destinations.

Routing is performed by the PostGIS extension pgrouting using the OpenStreetMap road network, 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 Docker containers deployed on a cloud computing service, such as Amazon Web Services (AWS). There is one job for each county. The travel network for each job covers the county and a 100 km buffer around it. All origins within the county are routed to all destinations within the county and its 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 travel times from each tract in the country to every destination 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, daytime trips in large cities such as New York and Philadelphia (see Appendix A).

The pipeline also computes on public 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). As for driving, OTP calculations can be containerized and distributed across many compute nodes. This method is relevant only where significant public transportation networks exist and where (GTFS) scheduling data are available. We focus on the 40 transit systems in the United States with the highest ridership. Running in parallel, travel times on these systems can be calculated in a few hours.

This pipeline significantly reduces the cost of calculating very-large OD matrices. A matrix of this size and granularity would cost over half a million dollars through the Google Maps API, and the license would not allow caching this result. A single term license of ArcGIS NetworkAnalyst 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.

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 will stimulate their use in this context. Nevertheless, though our Docker containers are publicly available, setting up the cloud computing may be a significant undertaking for a small project. We have therefore taken pains to ensure the outputs – travel time matrices – are also easily available for other researchers.

4. Access Results

The modelled accessibility of primary care is shown in Figure 1. Results are shown as a fractional deviation from the national mean, $\overline{\text{RAAM}}$. In other words, it is the spatial access ratio (SPAR) minus one: $\text{RAAM}(r)/\overline{\text{RAAM}} - 1$. A value of 0.5 corresponds to costs 50% higher than the national mean. 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 2A shows that the results are consistent over a broad range of values of τ. That Figure shows the Spearman's rank correlation coefficient for results derived with different values of τ, ranging from 15 minutes to 3 hours. Compared to the default setting of $\tau = 60$ minutes, those correlations exceed 90 percent for τ between 20 minutes and $\tau = 3$ hours. The choice of τ affects the total (unit-less) cost by making long trips (even) less palatable and it affects regions' costs relative to the national mean. But it has little effect on the accessibility rank.

Figure 2B shows the same rank correlations, across alternative models. The rank correlations to the nominal model exceed 95%; this suggests that the baseline results are quite robust. More broadly, the consistency among the RAAM variants is very high, with rank correlations above 90% The sole exception, at "merely" 86%, is between the $\tau = 30$ minutes model and the $(t/\tau)^2$ variant. The $\tau = 30$ minute setting values travel costs very dear, whereas the squared costs model makes short trips almost costless. That model is slightly less consistent with all of the other models.

The central aim of this project is to identify locations with high access costs, especially in rural America. Figure 3 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 2A, this behavior is robust against different choices for τ.

The left panel of that Figure shows the same trends for the modified models: the non-linear functional forms for the patient response to travel time, and multiple 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, commute-length trips to 0. This is consistent with the deweighted temporal access costs seen for $\tau = 90$ minutes, in Figure 3. By contrast the model with rurality-dependent willingness to travel, τ_r, shows a compression between the most- and least-rural areas. In this case, the lower rural disaffinty for travel reduces costs both directly by lowering the "cost per minute," and indirectly, by making it cheaper to use less-congested locations.

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 4A-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 1). Past work has documented the many non-spatial barriers that these populations face, which may be more significant. [12, 48]

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 τ for RAAM. We follow the specifications of the initial papers. For E2SFCA, we use Gaussian weights and three distance decay bands of 10, 20, and 30 minutes. [18] 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 $\beta = 640$. [19] At the national level, the weighted Pearson correlation between E2SFCA and 3SFCA is higher (0.80) than between them and RAAM (-0.52 and -0.61). The respective Spearman correlations are 0.87, -0.76, and -0.86; the 3SFCA is about as strongly correlated with RAAM as with E2SFCA. The correlations for the floating catchment approaches to RAAM and its various modifications are included in Figure 2B. These correlations are weaker than those among the RAAM's variants, but still fairly high: above 65% for E2SFCA and above 75% for 3SFCA, with the default settings. The enhanced 2SFCA has just three distance bands (10, 20, and 30 minutes) instead of four (adding 60 minutes) in the 3SFCA. This results in lower correlations to RAAM. With a more-stringent time setting of $\tau = 30$ minutes, RAAM's correlation to E2SFCA rises above 0.8). Alternatively, the E2SFCA correlation to RAAM ($\tau = 60$ minutes) is also stronger, with wider time bands.

Figure 5 shows changes in quantile between RAAM and floating catchment methods, in bands of rurality. Positive quantile shifts denote that the accessibility rank calculated with RAAM is higher than through FCA approaches. The Figure shows that RAAM predicts better access in fringe metro regions than E2SFCA and 3SFCA, but worse relative access in rural areas. 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. With a broader distance decay, the E2SFCA fringe metro are much closer to RAAM.

Finally, Figure 6A 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 model shows substantially lower bias than the single-origin model. We again emphasize that the dual origin preference fraction requires an assumption: that patients who seek care at work locations outside

their home region receive that care outside of that home region. The floating catchment results are biased in the opposite direction of RAAM, towards too little care in the home region. The magnitude of this bias for E2SFCA lies between the single and dual-origin models; the 3SFCA has the lowest bias of the four models shown. Figure 6B shows the sum of these weighted, squared "errors" as a function of τ. This shows that the RAAM error falls with increasing τ; although the slope evens out at $\tau \gtrsim 60$ minutes, it does not reach a minimum. So how does RAAM fare? This simple comparison suggests that non-home activities should be taken into account, but that with this modification RAAM performs comparably with the FCA models. However, we emphasize that though the dual-origin model substantially reduces the bias on this auxiliary measure of localization, it scarcely changes the access results. The Spearman's correlation of the nominal model with this one is 0.97.

5. Discussion

The preceding Sections have presented RAAM, an intuitive and easily extensible model for calculating accessibility, as well as a technical strategy for quickly deriving the travel times that are necessary for this calculation and which would otherwise be quite expensive. We have aimed to demonstrate the breadth of realistic modifications that can be incorporated into the model. Many of the extensions have previously been included in floating catchment methods. Some have not: the basic, competitive interactions of the model cannot be incorporated in FCA methods, and the multiple-origin variant is new to health accessibility. Models that downweight short commutes (dual-origin or $(t/\tau)^2$) compress costs between the cores and fringes of major metro areas; models that allow for more-palatable travel per minute in rural areas reduce the urban/rural divide. But by contrast with the national variation in access, these effects are small. The internal consistency of results is very high and lend credence to our nominal findings. The easy manipulability of the model is a distinct advantage in probing results.

RAAM incorporates different assumptions about consumer choice and demand elasticity than baseline 2SFCA methods. On the other hand, the distance preference function of the 3SFCA results in less-elastic consumption. With the distance decay settings of the original papers, the calculated accessibility ranks from 3SFCA are more strongly correlated than the E2SFCA results with RAAM. These conclusions, and any regarding the relative accuracy of the FCA methods, depend on the distance decay and model parameters used.

RAAM tends to result in smoother access maps than and E2SFCA or 3SFCA. It is not unusual to find E2SFCA and 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 1 does have a few local outliers, but these tracts are all islands or have extremely inaccessible centroids (for instance, the woods south of International Falls, MN).

We have also suggested new methods and data sources, for confronting the models with data, using the "preference index" (patient localization) of the Primary Care Service Area file. The localization calculated with RAAM and FCA methods (with the used catchment sizes and functions) have opposite biases: RAAM predicts too much care at home while FCA underpredicts it. The dual-origin RAAM model reduces this positive bias, but the magnitude of its bias remains larger than those in 3SFCA results. These results demand reiteration of several caveats. (a) Lower bias in localization does not necessarily correspond to large changes in accessibility. The localization bias is reduced by a third from single to dual-origin RAAM, but their Spearman's rank correlation is over 95%. (b) The Medicare Part B population used to derive these fractions may have different levels of home-region care than the general population. (c) The models do not have to place demand at the correct location to reduce bias in these comparisons. They need only avoid the home region the correct amount. This "not-here" structure is the essence of the gravity model. Better data are needed to understand and model patient choices, and the actual accessibility that they face.

Our technical work on travel matrices allows a tract-level calculation for the full United States. We see this data product as an important contribution. Measuring accessibility is not an end in and of itself. It is a result of socioeconomic conditions that demand continued study, and a potential

driver of health conditions in the population. But despite widespread adoption within geography of more-sophisticated models of accessibility, simplistic provider ratios are ubiquitous in health research. Geographers should work to share better measures.

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, per common practice of recent work. This procedure usually yields reasonable locations, but 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 suppy 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. [28] 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). This is related to the point made earlier: RAAM ignores both the intensive and extensive consumption response to prices. We look forward to modelling this demand elasticity, in coming work.

6. Conclusions

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 fundamentally different from floating catchment methods. RAAM is a versatile framework. It is trivial to alter patient responses to distance in FCA methods but RAAM also allows for multiple patient origin locations or travel modes – modifications that are unavailable in floating catchment methods, or have previously been demonstrated only at fairly small scale (regions or cities).

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. 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.

References

[1] Starfield, B., "Is primary care essential?" *The Lancet,* **344**:8930 1129-1133 (1994). DOI: 10.1016/S0140-6736(94)90634-3. Originally published as Volume 2, Issue 8930

[2] Shi, L., Starfield, B., Kennedy, B., and Kawachi, I., "Income inequality, primary care, and health indicators," *Journal of Family Practice,* **48**:4 275-284 (April, 1999).

[3] Shi, L., Macinko, J., Starfield, B., Politzer, R., Wulu, J., and Xu, J., "Primary care, social inequalities and all-cause, heart disease and cancer mortality in US counties: a comparison between urban and non-urban areas," *Public Health,* **119**:8 699-710 (2005a). DOI: 10.1016/j.puhe.2004.12.007.

[4] Shi, L., Macinko, J., Starfield, B., Politzer, R., Wulu, J., and Xu, J., "Primary Care, Social Inequalities, and All-Cause, Heart Disease, and Cancer Mortality in US Counties, 1990," *American Journal of Public Health,* **95**:4 674-680 (2005b). DOI: 10.2105/AJPH.2003.031716.

[5] Kringos, D. S., Boerma, W., Zee, J., and Groenewegen, P., "Europes Strong Primary Care Systems Are Linked To Better Population Health But Also To Higher Health Spending," *Health Affairs,* **32**:4 686-694 (2013). DOI: 10.1377/hlthaff.2012.1242.

[6] Starfield, B., Shi, L., and Macinko, J., "Contribution of Primary Care to Health Systems and Health," *The Milbank Quarterly,* **83**:3 457-502 (2005). DOI: 10.1111/j.1468-0009.2005.00409.x.

[7] Shi, L., "The Impact of Primary Care: A Focused Review," *Scientifica,* **2012** (2012). DOI: 10.6064/2012/432892.

[8] Hart, L. G., Salsberg, E., Phillips, D. M., and Lishner, D. M., "Rural Health Care Providers in the United States," *The Journal of Rural Health,* **18**:5 211-231 (2002). DOI: 10.1111/j.1748-0361.2002.tb00932.x.

[9] Weinhold, I. and Gurtner, S., "Understanding shortages of sufficient health care in rural areas," *Health Policy,* **118**:2 201-214 (November, 2014). DOI: 10.1016/j.healthpol.2014.07.018.

[10] Khan, A. A. and Bhardwaj, S. M., "Access to Health Care: A Conceptual Framework and its Relevance to Health Care Planning," *Evaluation & the Health Professions,* **17**:1 60-76 (1994). DOI: 10.1177/016327879401700104.

[11] Penchansky, R. and Thomas, J. W., "The Concept of Access: Definition and Relationship to Consumer Satisfaction," *Medical Care,* **19**:2 127-140 (1981). DOI: 10.2307/3764310.

[12] Aday, L. A. and Andersen, R., "A Framework for the Study of Access to Medical Care.," *Health Services Research,* **9**:3 208-220 (1974).

[13] Andersen, R. M., McCutcheon, A., Aday, L. A., Chiu, G. Y., and Bell, R., "Exploring dimensions of access to medical care," *Health Services Research,* **18**:1 49-74 (1983).

[14] Kringos, D. S., Boerma, W. G., Hutchinson, A., Zee, J., and Groenewegen, P. P., "The breadth of primary care: a systematic literature review of its core dimensions," *BMC Health Services Research,* **10**:1 65 (Mar 13, 2010). DOI: 10.1186/1472-6963-10-65.

[15] Luo, W. and Wang, F., "Measures of Spatial Accessibility to Health Care in a GIS Environment: Synthesis and a Case Study in the Chicago Region," *Environment and Planning B: Planning and Design,* **30**:6 865-884 (2003). DOI: 10.1068/b29120.

[16] Luo, W., "Using a GIS-based floating catchment method to assess areas with shortage of physicians," *Health & Place,* **10**:1 1-11 (2004). DOI: 10.1016/S1353-8292(02)00067-9.

[17] McGrail, M. R. and Humphreys, J. S., "Measuring spatial accessibility to primary care in rural areas: Improving the effectiveness of the two-step floating catchment area method," *Applied Geography,* **29**:4 533-541 (2009). DOI: 10.1016/j.apgeog.2008.12.003.

[18] Luo, W. and Qi, Y., "An enhanced two-step floating catchment area (E2SFCA) method for measuring spatial accessibility to primary care physicians," *Health & Place,* **15**:4 1100-1107 (2009). DOI: 10.1016/j.healthplace.2009.06.002.

[19] Wan, N., Zou, B., and Sternberg, T., "A three-step floating catchment area method for analyzing spatial access to health services," *International Journal of Geographical Information Science,* **26**:6 1073-1089 (2012). DOI: 10.1080/13658816.2011.624987.

[20] Li, Z., Serban, N., and Swann, J. L., "An optimization framework for measuring spatial access over healthcare networks," *BMC Health Services Research,* **15**:1 273 (Jul 17, 2015). DOI: 10.1186/s12913-015-0919-8.

[21] Gentili, M., Harati, P., Serban, N., O'Connor, J., and Swann, J., "Quantifying Disparities in Accessibility and Availability of Pediatric Primary Care across Multiple States with Implications for Targeted Interventions," *Health Services Research,* **53**:3 1458-1477 (2017). DOI: 10.1111/1475-6773.12722.

[22] Mao, L. and Nekorchuk, D., "Measuring spatial accessibility to healthcare for populations with multiple transportation modes," *Health & Place,* **24** 115-122 (2013). DOI: 10.1016/j.healthplace.2013.08.008.

[23] Langford, M., Higgs, G., and Fry, R., "Multi-modal two-step floating catchment area analysis of primary health care accessibility," *Health & Place,* **38** 70-81 (2016). DOI: 10.1016/j.healthplace.2015.11.007.

[24] Tao, Z., Yao, Z., Kong, H., Duan, F., and Li, G., "Spatial accessibility to healthcare services in Shenzhen, China: improving the multi-modal two-step floating catchment area method by estimating travel time via online map APIs," *BMC Health Services Research,* **18**:1 345 (May 09, 2018). DOI: 10.1186/s12913-018-3132-8.

[25] Lin, Y., Wan, N., Sheets, S., Gong, X., and Davies, A., "A multi-modal relative spatial access assessment approach to measure spatial accessibility to primary care providers," *International Journal of Health Geographics,* **17**:1 33 (Aug 23, 2018). DOI: 10.1186/s12942-018-0153-9.

[26] Fransen, K., Neutens, T., Maeyer, P. D., and Deruyter, G., "A commuter-based two-step floating catchment area method for measuring spatial accessibility of daycare centers," *Health & Place,* **32** 65-73 (2015). DOI: 10.1016/j.healthplace.2015.01.002.

[27] McGrail, M. R. and Humphreys, J. S., "Spatial access disparities to primary health care in rural and remote Australia," *Geospatial Health,* **10**:2 138-143 (2015). DOI: 10.4081/gh.2015.358.

[28] Ricketts, T. C., Goldsmith, L. J., Holmes, G. M., Randolph, R., Taylor, J., and Ostermann, J., "Designating Places and Populations as Medically Underserved: A Proposal for a New Approach," *Journal of Health Care for the Poor and Underserved,* **18**:3 567-589 (August, 2007). DOI: 10.1353/hpu.2007.0065.

[29] Goodman, D. C., Mick, S. S., Bott, D., Stukel, T., Chang, C.-h., Marth, N., *et al.*, "Primary Care Service Areas: A New Tool for the Evaluation of Primary Care Services," *Health Services Research,* **38**:1p1 287-309 (2003). DOI: 10.1111/1475-6773.00116.

[30] Radke, J. and Mu, L., "Spatial Decompositions, Modeling and Mapping Services Regions to Predict Access to Social Programs," *Geographic Information Science,* **6**:2 105-112 (December, 2000). DOI: 10.1080/10824000009480538.

[31] Wang, F. and Luo, W., "Assessing spatial and nonspatial factors for healthcare access: towards an integrated approach to defining health professional shortage areas," *Health & Place,* **11**:2 131-146 (2005). DOI: 10.1016/j.healthplace.2004.02.003. Special section: Geographies of Intellectual Disability

[32] Joseph, A. E. and Bantock, P. R., "Measuring potential physical accessibility to general practitioners in rural areas: A method and case study," *Social Science & Medicine,* **16**:1 85-90 (1982). DOI: 10.1016/0277-9536(82)90428-2.

[33] Huff, D. L., "A Probabilistic Analysis of Shopping Center Trade Areas," *Land Economics,* **39**:1 81-90 (1963). DOI: 10.2307/3144521.

[34] Wang, F., "Inverted Two-Step Floating Catchment Area Method for Measuring Facility Crowdedness," *The Professional Geographer,* **70**:2 251-260 (2018). DOI: 10.1080/00330124.2017.1365308.

[35] Guagliardo, M. F., "Spatial accessibility of primary care: concepts, methods and challenges," *International Journal of Health Geographics,* **3**:1 3 (Feb 26, 2004). DOI: 10.1186/1476-072X-3-3.

[36] McGrail, M. R., "Spatial accessibility of primary health care utilising the two step floating catchment area method: an assessment of recent improvements," *International Journal of Health Geographics,* **11**:1 50 (Nov 16, 2012). DOI: 10.1186/1476-072X-11-50.

[37] Yasaitis, L. C., Bynum, J. P. W., and Skinner, J. S., "Association Between Physician Supply, Local Practice Norms, and Outpatient Visit Rates," *Medical Care,* **51**:6 524-531 (2013). DOI: 10.1097/MLR.0b013e3182928f67.

[38] Wan, N., Zhan, F. B., Zou, B., and Chow, E., "A relative spatial access assessment approach for analyzing potential spatial access to colorectal cancer services in Texas," *Applied Geography,* **32**:2 291-299 (2012). DOI: 10.1016/j.apgeog.2011.05.001.

[39] Donohoe, J., Marshall, V., Tan, X., Camacho, F. T., Anderson, R., and Balkrishnan, R., "Evaluating and comparing methods for measuring spatial access to mammography centers in Appalachia," *Health Services and Outcomes Research Methodology,* **16**:1 22-40 (Jun 01, 2016). DOI: 10.1007/s10742-016-0143-y.

[40] Jia, P., Wang, F., and Xierali, I. M., "Using a Huff-Based Model to Delineate Hospital Service Areas," *The Professional Geographer,* **69**:4 522-530 (2017). DOI: 10.1080/00330124.2016.1266950.

[41] McGrail, M. R., Humphreys, J. S., and Ward, B., "Accessing doctors at times of need: measuring the distance tolerance of rural residents for health-related travel," *BMC Health Services Research,* **15**:212 (2015). DOI: 10.1186/s12913-015-0880-6.

[42] Hägerstraand, T., "What about people in regional science?" *Papers in Regional Science,* **24**:1 7-24 (1970). DOI: 10.1111/j.1435-5597.1970.tb01464.x.

[43] Horton, F. E. and Reynolds, D. R., "Action Space Formation: A Behavioral Approach to Predicting Urban Travel Behavior," *Highway Research Record,* **322** 136-148 (1970).

[44] Patterson, Z. and Farber, S., "Potential Path Areas and Activity Spaces in Application: A Review," *Transport Reviews,* **35**:6 679-700 (2015). DOI: 10.1080/01441647.2015.1042944.

[45] Donabedian, A., "Models for Organizing the Delivery of Personal Health Services and Criteria for Evaluating Them," *The Milbank Memorial Fund Quarterly,* **50**:4 103-154 (1972). DOI: 10.2307/3349436.

[46] Graham, M. R., Kutzbach, M. J., and McKenzie, B., *Design Comparison of LODES and ACS Commuting Data Products.* Technical Report 14-38 (Center for Economic Studies, U.S. Census Bureau; October 2014).

[47] OpenStreetMap contributors, "North America data retrieved from http://download.geofabrik.de/." Accessed June 1, 2018.

[48] Grumbach, K., Vranizan, K., and Bindman, A. B., "Physician Supply And Access To Care In Urban Communities," *Health Affairs,* **16**:1 71-86 (1997). DOI: 10.1377/hlthaff.16.1.71.

[49] Ruggles, S., Genadek, K., Goeken, R., Grover, J., and Sobek, M., "Integrated Public Use Microdata Series," *University of Minnesota,* **7.0** (2017). DOI: 10.18128/D010.V7.0..

[50] McGrail, M., Wingrove, P. M., Petterson, S. M., Humphreys, J., Russell, D., and Bazemore, A. W., "Measuring the attractiveness of rural communities in accounting for differences of rural primary care workforce supply," *Rural and Remote Health,* **17**:3925 (April, 2017). DOI: 10.22605/RRH3925.

Appendix A. Comparison of travel times with commercial suppliers.

Figure 7 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, $\overline{\ell}$, to the minimum cost location available to them, $\underline{\ell}$. 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 ($\not r$): $\sum_{r'} d_{r'\ell} = d_{r\ell} + d_{\not r\ell}$. Residents can alter their own allocations $d_{r\ell}$ but they cannot affect others' location decisions; $d_{\not r\underline{\ell}}$ and $d_{\not r\overline{\ell}}$ are fixed in each iteration. The physician supply of a location s_ℓ and the times necessary to reach it $t_{r\ell}$ are permanently immutable.

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

$$\frac{\left(d_{r\underline{\ell}} + d_{\not r\underline{\ell}}\right)/s_{\underline{\ell}}}{\rho} + \frac{t_{r\underline{\ell}}}{\tau} = \frac{\left(d_{r\overline{\ell}} + d_{\not r\overline{\ell}}\right)/s_{\overline{\ell}}}{\rho} + \frac{t_{r\overline{\ell}}}{\tau}$$

$$\frac{d_{r\underline{\ell}} + d_{\not r\underline{\ell}}}{s_{\underline{\ell}}} = \left(t_{r\overline{\ell}} - t_{r\underline{\ell}}\right)\frac{\rho}{\tau} + \frac{d_{r\overline{\ell}} + d_{\not r\overline{\ell}}}{s_{\overline{\ell}}}.$$

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

$$\frac{d_{r\underline{\ell}} + d_{\not r\underline{\ell}}}{s_{\underline{\ell}}} = \left(t_{r\overline{\ell}} - t_{r\underline{\ell}}\right)\frac{\rho}{\tau} + \frac{d_r - d_{r\underline{\ell}} + d_{\not r\overline{\ell}}}{s_{\overline{\ell}}}.$$

Now isolate $d_{r\underline{\ell}}$ on one side:

$$d_{r\underline{\ell}} = \left(\frac{s_{\underline{\ell}} \times s_{\overline{\ell}}}{s_{\underline{\ell}} + s_{\overline{\ell}}}\right)\left[\left(t_{r\overline{\ell}} - t_{r\underline{\ell}}\right)\frac{\rho}{\tau} + \frac{d_r + d_{\not r\overline{\ell}}}{s_{\overline{\ell}}} - \frac{d_{\not r\underline{\ell}}}{s_{\underline{\ell}}}\right].$$

This represents the values of $d_{r\underline{\ell}}$ (and $d_{r\overline{\ell}} = d_r - d_{r\underline{\ell}}$) that equalize costs between the $\underline{\ell}$ and $\overline{\ell}$. 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 $\mathbf{1}(R_i = R_j)$ 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 $\mathcal{W}$, 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:

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

Appendix D. The Determinants of Access

The findings of the preceding sections illustrate the well-known shortages of primary care in rural America. [8, 9] Region by region, cities have lower access costs for care than their hinterlands. Our modified models show fairly consistent results for rural populations. But what drives poor access?

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,[5] 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, each of which contrasts 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 (τ = 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 is drawn not from the PCSA/AMA file but from 2010 ACS public use microdata. [49] This records physicians at home rather than at work, and it therefore includes all physicians – not just primary care providers. 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. Isolated, 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.[6] 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: $R^2 = 0.49$ in tandem with population density, or $R^2 = 0.28$ on its own. Constraining the sample to the five hundred least-dense

[5] 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 $R^2 = 0.37$.

[6] 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.

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 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 appears almost 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. This point has been made recently by McGrail and colleagues. [50] There is an important distinction between policies geared towards rural doctors and those geared towards underserved populations.

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.

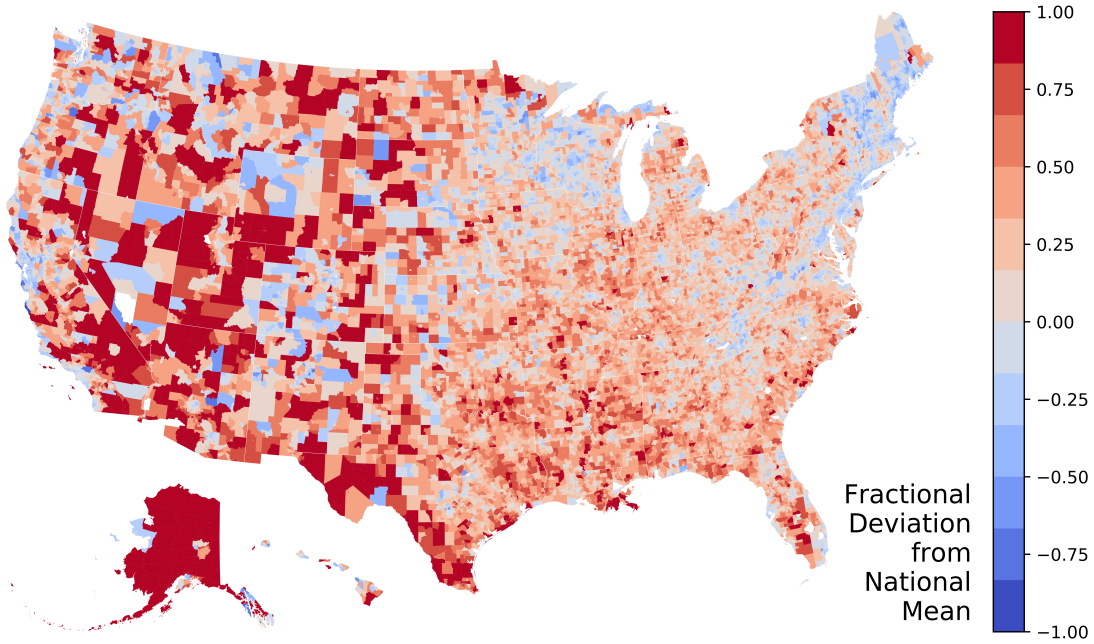

FIGURE 2. Rank correlations of access costs across Census tracts, for (A) parameter settings and (B) alternative models. Panel (A) shows that varying the trade-off τ between travel time and congestion weight from $\tau = 20$ minutes to 3 hours, the rank correlations to results from our baseline setting of $\tau = 60$ minutes are over 0.9. Panel (B) displays modifications of RAAM, and comparisons to floating catchment results. The internal consistency of RAAM is mostly above 90% Larger differences are apparent with respect to the FCA methods, especially E2SFCA.

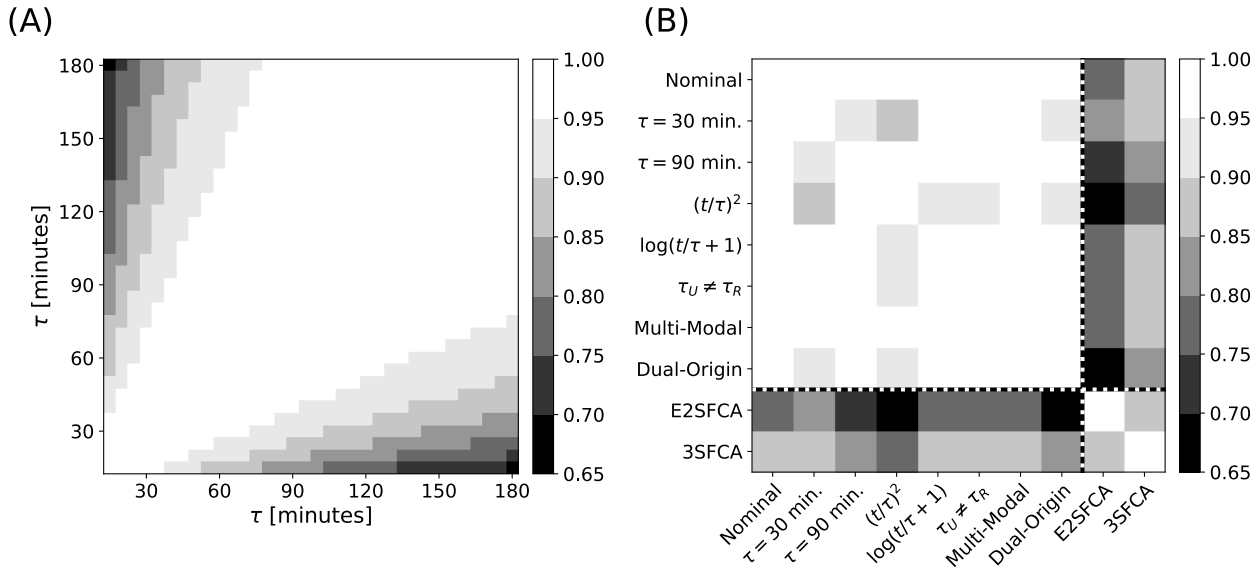

FIGURE 3. Spatial access by NCHS rurality, for varying definitions of τ and modified models. The boxes show the median values and interquartile ranges per rurality level, and the whiskers show then 10th and 90th percentiles. After demeaning and normalizing each distribution (taking the z-scores), the different values of τ show markedly consistent trends in access against rurality. The dual-origin and squared-costs models reduce differences between metropolitan areas and their fringes, while the stronger urban disaffinity for travel ($\tau_U < \tau_R$) leads to a compression of costs between urban and rural regions, with respect to nominal.

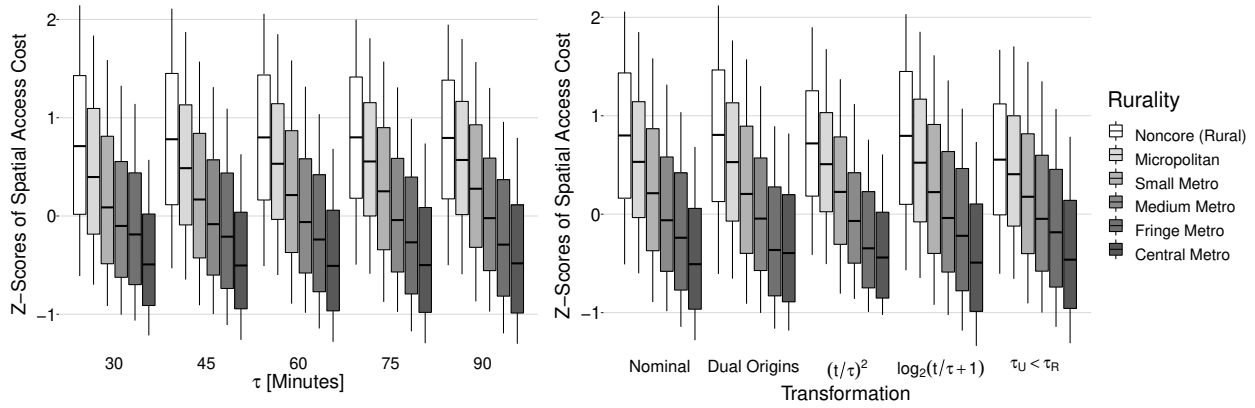

FIGURE 4. Differences in calculated access costs, between driving-based or multimodal travel, or between driving and public transportation only, in Cook County, Illinois. Travel times are longer on public transportation than by car, leading to higher costs for populations using transit (A). This leads 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).

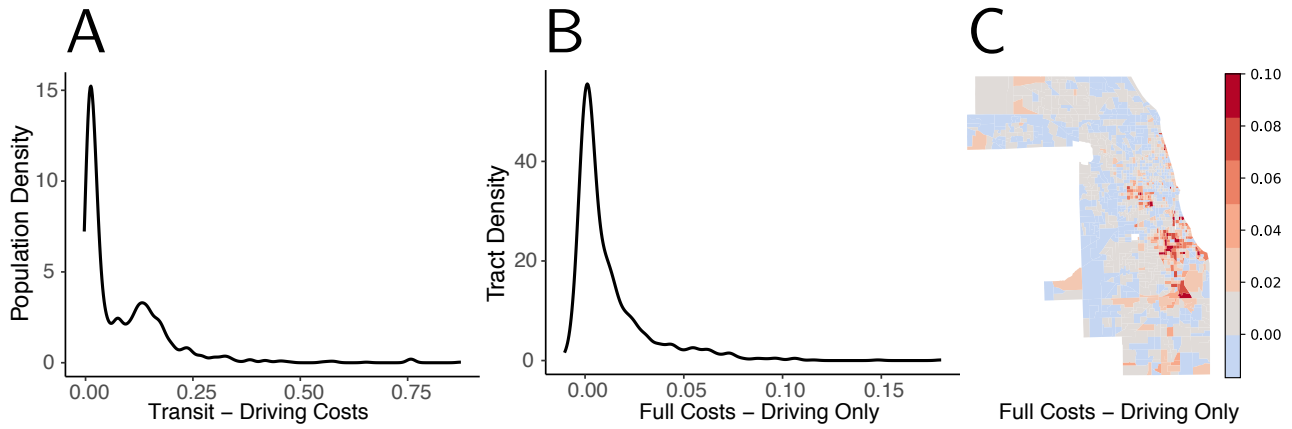

FIGURE 5. The difference in quantiles of for accessibility costs versus (negative) two- and three-stage floating catchment accessibility, for population-weighted Census tracts. Positive differences in large metro counties denote better accessibility rank according to RAAM, while negative values in micropolitan and rural counties denote worse accessibility. 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 central time bands of 10, 20, and 30 minutes. For the 3SFCA we also include the larger band at 60 minutes, and we set $\beta = 640$. [18, 19].

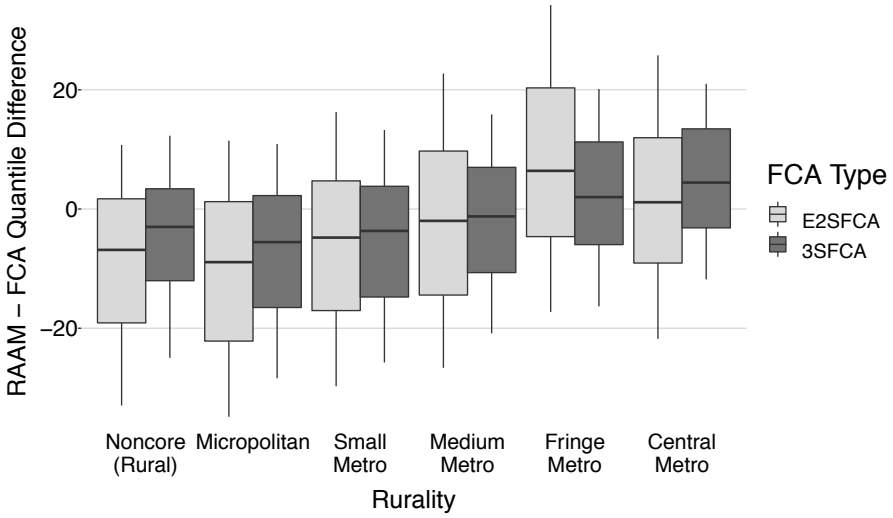

FIGURE 6. Difference between use modeled in Primary Care Service Areas, and the "Preference Fractions" (individual-weighted use) measured by the Dartmouth Institute through Medicare claims (panel A). The difference is significantly reduced by the use of a dual-origin model but the bias is not eliminated. The population-weighted sums of squared errors (panel B) do not reach minima with respect to τ: they remain biased high.

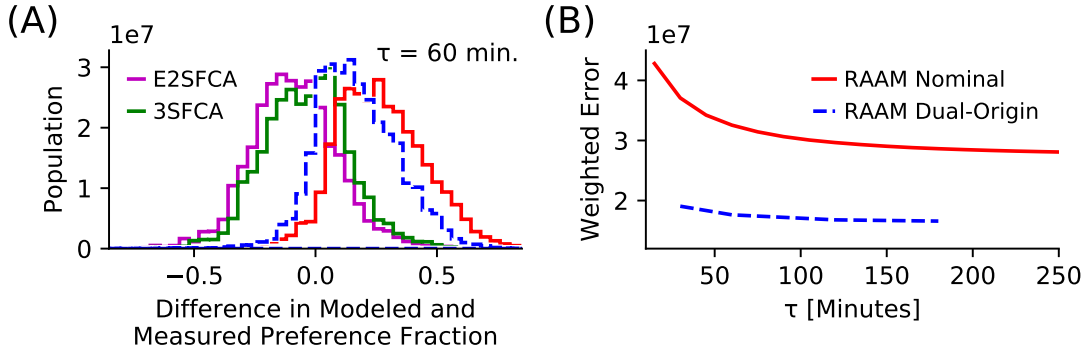

FIGURE 7. Driving travel times from our distributed pipeline are shown against those from commercial providers, for five major cities and five (random) rural counties.

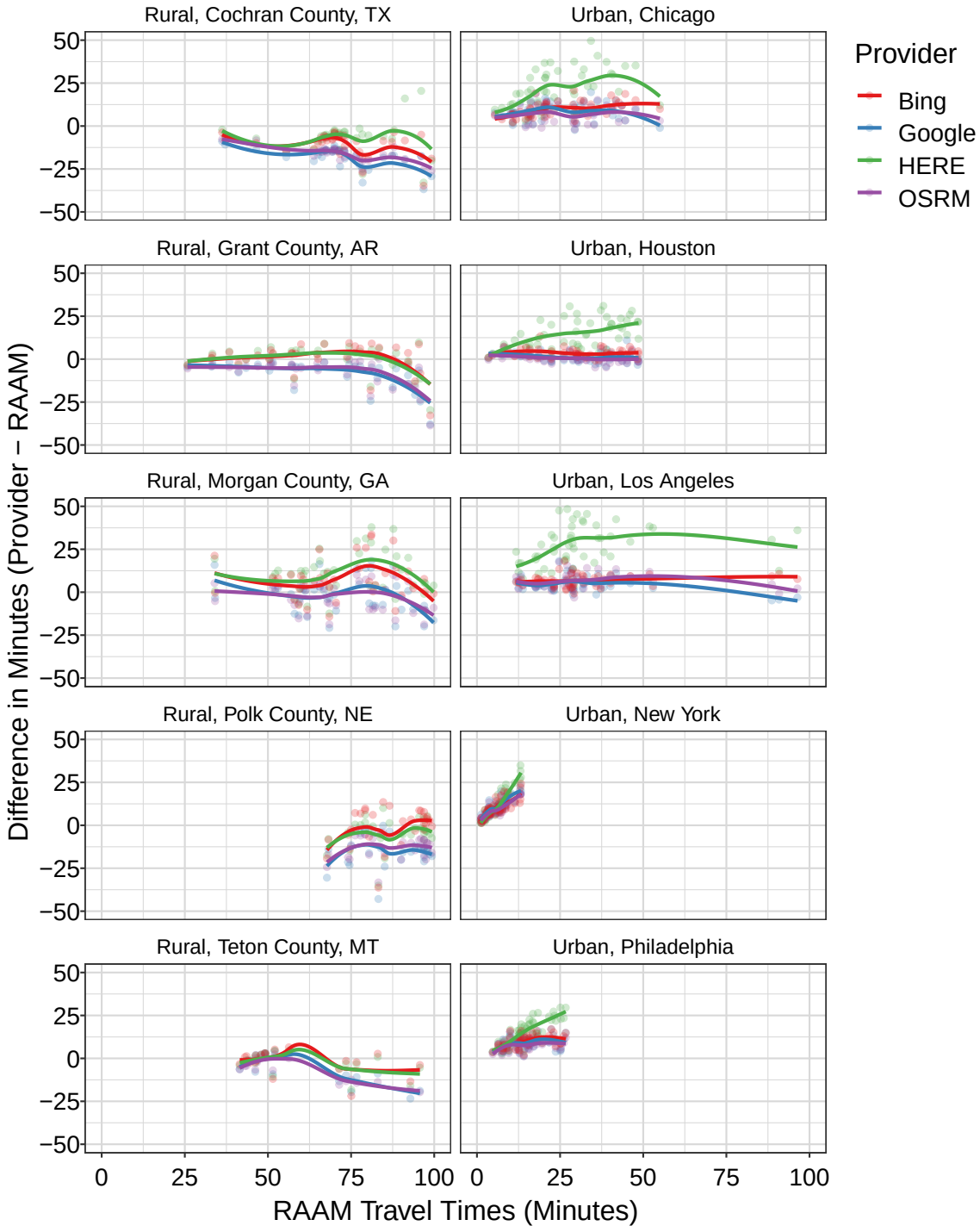

TABLE 1. Accessibility costs, congestion costs, and doctor residential doctors are regressed on population density and socioeconomic characteristics, at the PUMA level. Educational attainment is strongly predictive of accessibility of lower primary care costs. After removing travel costs (middle panel), it is equally predictive as population density. Focusing on physicians' residential location decisions, regional education is far more predictive than density.

	Total RAAM Cost				Congestion Cost			Physicians / 1k Residents		
	1	2	3	4	1	2	3	1	2	3
Intercept	2.06*	1.27*	2.02*	-1.79*	1.95*	1.32*	1.90*	-3.16*	-1.98*	-2.09*
	(0.03)	(0.01)	(0.03)	(0.29)	(0.03)	(0.01)	(0.03)	(0.48)	(0.11)	(0.34)
Log Population	-0.07*		-0.06*	-0.07*	-0.06*		-0.04*	0.40*		0.01
Density	(0.00)		(0.00)	(0.00)	(0.00)		(0.00)	(0.03)		(0.02)
Bachelor's		-0.96*	-0.65*	-0.85*		-0.91*	-0.66*		16.97*	16.93*
		(0.03)	(0.03)	(0.04)		(0.03)	(0.03)		(0.35)	(0.38)
Log Median				0.36*						
HH Income				(0.03)						
Poverty				1.36*						
				(0.13)						
Black				-0.05						
				(0.03)						
R^2	0.37	0.28	0.49	0.53	0.26	0.25	0.37	0.07	0.53	0.53
N	2071	2071	2071	2071	2071	2071	2071	2071	2071	2071

Standard errors in parentheses; * $p < 0.001$.