INTERNATIONAL JOURNAL OF INTELLIGENT CONTROL AND SYSTEMS
VOL. 14, NO. 1, MARCH 2009, 51-61

A GIS Aware Agent-Based Model of Pathogen Transmission

Ryan C. KENNEDY, Kelly E. LANE, S. M. Niaz ARIFIN, Agustín FUENTES, Hope HOLLOCHER, and Gregory R. MADEY

***Abstract*—Agent-based modeling (ABM) is quickly becoming a choice tool for researchers, as it is very adept at modeling natural phenomena. Spatial data has the ability to further increase the usefulness of an ABM. One way that this can be achieved is through the inclusion of geographical information system (GIS) data. Such integration between ABM and GIS may sound trivial, but it is difficult to do effectively, particularly as the complexity of GIS data and the amount of agents increase. Here, we present methods and recommendations on including GIS data in a complex simulation model. We demonstrate our techniques on an advanced epidemiological model.**

***Index Terms*—performance, scientific simulation, raster data, vector data, epidemiological modeling**

1. Introduction

SIMULATIONS of real-world phenomena have the potential to be valuable to researchers. Rather than relying on complex, approximate equations, agent-based models (ABMs) rely on more natural behavioral rules [16]. This leads to a more direct translation from natural phenomena to a simulation model. It is logical to integrate spatial data into the simulation environment; however, as Gilbert [14] pointed out, utilizing geographical information system (GIS) data for dynamic agents is a difficult challenge that has not yet been adequately solved. Although GIS data has successfully been integrated into ABMs for several years, the ability to run complex simulations with thousands of GIS aware agents is computationally challenging. In this article, we present several methods of integrating GIS data into a simulation environment. We describe an epidemiological model that utilizes GIS data and offer insight on how to efficiently integrate GIS data into a model, depending on the model's complexity and needs.

The organization of this paper is as follows. In section 2, we discuss the integration of ABMs and GIS data. Section 3 details our simulation model and section 4 provides a discussion. We finish with conclusions in section 5, followed by an acknowledgment and references.

2. Geographic Information System Data and Agent-based Simulations

GIS data has a variety of applications and spans many fields. Simply, a GIS is a system in which real-world environmental data is represented. Examples include rivers, governmental boundaries, rainfall, temperature, population distribution, and disease prevalence, among many others. GIS data is typically stored in raster or vector format. Raster data is characterized as a collection of pixels, or cells. These cells typically make up a grid-like structure, with each cell having its own attributes and properties. Vector data is coordinate-based; namely, data is represented by points, lines, and polygons. These features also have associated characteristics. While raster data lends itself directly to the grid-like frameworks of ABMs, it is subject to spatial resolution issues and requires large amounts of storage space. Vector data is more realistic, as it suffers less resolution loss, and is more easily stored. However, querying vector data can be very computationally expensive. For example, querying a set of complex polygons representing forests in an environment would require multiple, expensive queries to each polygon for each agent, unless some sort of indexing was performed. Figure 1 visually compares raster and vector data. A means to combine the benefits of raster and vector data to create GIS aware agents would be an important step in the advancement of ABMs with GIS data.

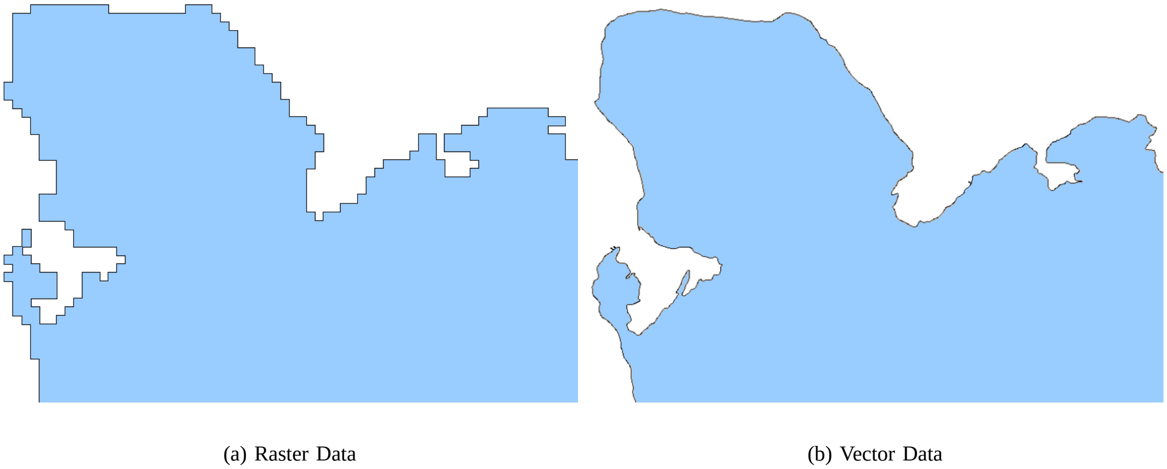
While previous studies have described ABMs coupled with GIS data, most existing models do not have agents that intelligently move based on their current environment. Castle, et al. [5] mention numerous toolkits and applications for this yet fail to go beyond the incorporation of GIS data into a model and into the realm of its effective use. Crooks [6] more deeply describes the realm of space within ABM and offers example applications but does not specifically address the underlying issue of how agents can most efficiently access GIS data. Anwar, et al. [1] describe a model built upon GIS data, but one that does not directly query it. Some models imply space, such as NOSOSIM [27], but few dynamically interact with GIS data. Gimblett [15], Keeling, et al. [18], and Brown et al. [4] describe aspects of the integration of ABMs and GIS data, but do not go into detail regarding approaches to efficiently create GIS aware agents. Moreover, standard means of linking agents with GIS data are computationally expensive and therefore not feasible for complex, large-scale simulation models. In many cases, only particular parts of a GIS are necessary for an ABM; utilizing a feature-rich GIS toolkit at simulation runtime is not typically advisable. We next describe the problem at hand and offer increasingly better solutions.

2.1. GIS Data in Agent-based Models

In traditional ABMs, agents typically move about a grid-like structure. GIS aware agents move about the same structure,

Manuscript received February 1, 2009. This work was supported in part by National Science Foundation under Grant No. BSC-0629787. This paper is extended from Kennedy et al. [20].

R.C. Kennedy, S. M. Niaz Arifin, and G.R. Madey are with the Department of Computer Science and Engineering. K.E. Lane and H. Hollocher are with the Department of Biological Sciences. A. Fuentes is with the Department of Anthropology. All authors are at the University of Notre Dame.


(a) Raster Data

(b) Vector Data

Fig. 1. Panels (a) and (b) show the northwest corner of Bali, Indonesia as represented by a raster and a vector file. Vector data is generally more precise than raster data.

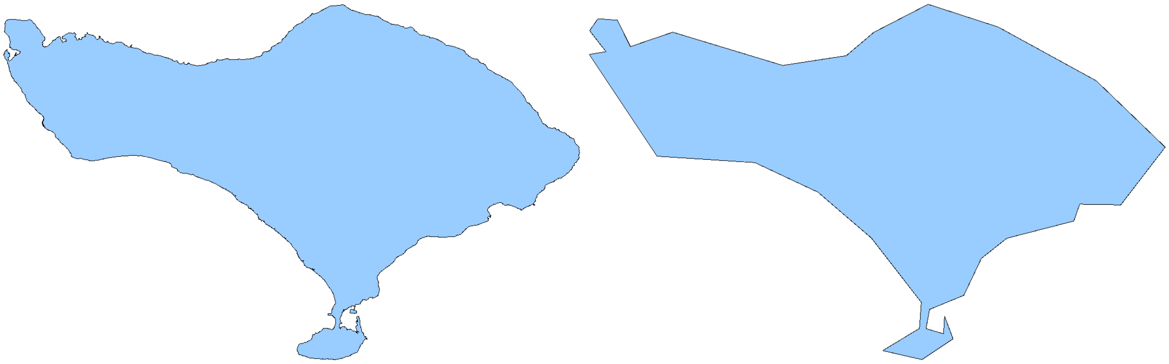
but in a manner such that each move is influenced by the surrounding environment, including nearby agents. A simple example would be allowing agents to move preferentially into one landscape over another. When an ABM environment is built upon GIS data, queries can be expensive, particularly with complex data or movement. As a general rule, the more complex the GIS data, the more difficult it is to efficiently utilize it within an ABM. Additionally, the more GIS data that is available, such as multiple landscape features, the more time-consuming it will be for agents to query. Put simply, at each timestep, an agent needs to query its unknown surroundings and make a decision regarding its next move. The more GIS data there is, the longer this will take. A common solution is to approximate GIS data to the level of granularity required for a given model. As such, the amount of GIS data is decreased while the integrity of the data required is maintained. We next describe several ways to access GIS data from a simulation, offering advantages and disadvantages for each.

2.1.1) Raster Queries: Raster-based spatial queries made through a spatial package can be costly, as the mechanisms by which agents access this data are typically not optimized for use in simulations. Additionally, storing and loading potentially large raster data files is inefficient at simulation runtime, particularly when not all of the data is necessary. Raster files are also not ideal for representing complex GIS data where fine-scale granularity is required. An advantage of utilizing raster data in an ABM is that it easily maps to traditional ABM grid spaces.

2.1.2) Spatial Queries: Spatial queries on vector-based GIS data are the most accurate way an agent can interact with GIS data. Here, an agent simply performs mathematical-based queries on the loaded GIS data to determine its surroundings. While very accurate, the cost of performing a spatial query increases as the complexity of the data increases. For example, it may be mathematically simple to query a rectangle to see whether an agent is contained within it; however, it is very mathematically expensive to do the same query on a large polygon. Repeatedly performing such queries is expensive, and this problem is exaggerated as the number of agents and the amount of spatial data increases. While indexing spatial data alleviates some redundancy, queries are still expensive.

2.1.3) Simplified Spatial Queries: The performance of spatial queries can be improved if the vector data is approximated in a manner such that the number of vertices in a line or polygon is decreased, while maintaining an appropriate level of data integrity. The Douglas-Peucker algorithm [8] is commonly used to perform such simplifications. This technique offers a speedup over traditional spatial queries, but at a cost of less accurate spatial data. However, repeatedly performing similar or identical spatial queries is redundant and can be remedied. Figure 2 shows a near 100% data simplification that maintains considerable data integrity.

2.1.4) Precalculated Query Matrix: Recognizing the drawbacks of earlier techniques, we developed a technique we call the precalculated query matrix. This technique relies on the advantages of raster data while utilizing the accuracy of vector data. Here, vector files are used in conjunction with spatial queries to build arrays of spatial data. Specifically, we iterate through the vector data, at a specified granularity, and perform spatial queries at each point, saving the results. This process is shown in Algorithm 1 and is performed for all available spatial data. The runtime for Algorithm 1 is *O(xyl)*, where *x* and *y* are the number of latitude and longitude values and *l* is the number of matrices. While time consuming, the expensive queries only need to be performed once for a given granularity, prior to simulation runtime. We utilize serialization to load the arrays into the simulation and agents can access the data in constant time. The main disadvantage to this method is that arrays of finer granularities will take longer to build, resulting in larger arrays and slightly longer query times. The advantages include agents that can more quickly query their environment and a simulation that scales well, both in terms of the amount of GIS data available and in the number of agents. Researchers also have the advantage of choosing a granularity to fit their needs. Currently, we use multiple precalculated query matrices.



(a) 10,000 Data Points

(b) 100 Data Points

Fig. 2. Panels (a) and (b) represent Bali, Indonesia with approximately 10,000 and 100 data points, respectively. Here, we reduce the number of points by almost 100%, but still retain considerable data integrity.

Algorithm 1 BuildPrecalculatedQueryMatrix

```
Let $X$ be the set of latitude values
Let $Y$ be the set of longitude values
Let $L$ be the set of GIS layers
Let $M$ be the Precalculated Query Matrix for a layer
for all $x \in X$ do
    for all $y \in Y$ do
        for all $l \in L$ do
            $M_l(x, y) \leftarrow$ SpatialQuery($l, x, y$)
        end for
    end for
end for
```

2.2. *GIS Aware Agents*

Previously, we listed ways by which agents can query their environment. Once agents are able to adequately and efficiently survey their surroundings, they must be able to make use of that data to become spatially aware. Our agents make use of precalculated query matrices for movement decisions. To display this movement on the native vector data, we use hash tables to "map" the native GIS latitude and longitude points to our matrices, and vice versa. This mapping avoids repetitive calculations, while allowing agents to find their real-world coordinates with ease. This also assists in enabling agents to move with complex rules, which we next describe.

2.2.1) Movement: Adding movement to agents in a GIS-based environment is challenging. With raster data, agents must perform tedious queries through the GIS system to determine the surrounding landscape. Spatial queries are inefficient too, as the queries can be redundant and take considerable time. Utilizing precalculated query matrices enables us to create many agents with complex and realistic movements in rapid time.

In traditional ABM cellular automata spaces, agent behavior is based on a von Neumann or Moore neighborhood. Specifically, von Neumann neighborhoods describe the four cells immediately adjacent to the current cell in a traditional square grid. A Moore neighborhood extends this to the surrounding eight adjacent cells, including those diagonally adjacent. Performing spatial queries on such spaces would be tedious and inefficient, particularly if the neighborhood was extended beyond a Moore neighborhood.

In our model, spatial movement is based on a Moore neighborhood, with allowance for larger neighborhoods. To move intelligently, agents must know the landscape they are currently in as well as the surrounding landscape. To represent possible transitions from one cell to another, we use a matrix of probabilistic movement values. This table consists of values representing the likelihood that an agent would move from a given landscape to another. Calculations are performed for each of the cells in the Moore neighborhood. A directional bias is also added to the agents so they are more likely to continue in the same general direction. Once the values for the surrounding cells have been calculated, they are normalized. We then use probabilities to determine the next location for the agent, if it moves at all. These calculations are performed quickly, as the lookups for the surrounding cells can be performed in constant time, allowing for realistic movement among agents. Figure 3 shows a simplified version of our movement on an example grid and Algorithm 2 describes dispersed movement algorithmically (time-dependent on the number of possible new locations).

Intelligent agents can be classified as simple reflex, model-based reflex, goal-based reflex, utility-bases, or as learning [26]. Based on the movement decisions described previously, our agents could be classified as utility-based, but with a stochastic-based utility functions and decisions. This classification fits our agents because they make decisions based upon utility - they are more content in certain landscapes, and their contentment is determined by their previous location and current landscape. We next describe the simulation model we have developed in more detail.

3. A Simulation Model of Pathogen Transmission

We have created a model, named LiNK and further described in Lane [22], to aid in the understanding of pathogen transmission patterns. This model was designed to simulate

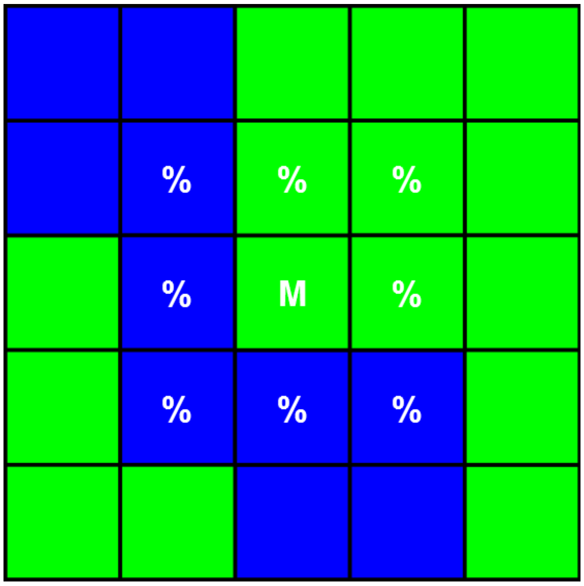


Fig. 3. Macaque Movement. The graphic above shows how a macaque *M* determines where to move in a landscape consisting of forests (green) and a river (blue). There are movement probabilities associated with landscape features. For example, a macaque would be more likely to enter a forest than a river. Here, we base movement on the immediate surrounding cells; however, it can be based on an arbitrary number of cells in an outward direction.

Algorithm 2 DISPERSEDMOVEMENT

```
Let $P$ be the probability a macaque moves to a given location $l$
Let $L_{t+1}$ be the set of possible locations for the next timestep
Let $l_{t+1}$ be the new location
Let $b_1$ be the directional bias
Let $b_2$ be the landscape bias
for each timestep $t$ do
  for all $l \in L_{t+1}$ do
    $l \leftarrow b_1 + b_2$
  end for
  $l_{t+1} \leftarrow$WEIGHTEDSELECTONADJUST($l \in L_{t+1}$)
end for
```

the spread of infection amongst long-tailed macaques (*Macaca fascicularis*, Figure 4) on the Indonesian island of Bali. We have coupled detailed GIS data with a deep knowledge of the macaque population to create a rich simulation.

3.1. Background

Several zoonotic diseases have recently emerged on the Asian landscape; macaques have been implicated as both hosts and reservoirs in these disease emergences in humans. Increasing anthropogenic landscape changes have increased the incidence of human to non-human primate interaction, potentially leading to bi-directional pathogen transmission events [7], [10], [22]. In our model, we evaluate how landscape changes might influence pathogen transmission patterns, based on the behavior and dispersal patterns of long-tailed macaques across the island of Bali. We specifically aim to address the following research questions:

Fig. 4. Female Macaque (*Macaca fascicularis*) and Infant. Photo courtesy of A. Fuentes.

1) What are potential rates and routes of pathogen transmission in macaques across the island?
2) How do pathogen life history parameters impact this transmission?
3) Do the answers change with the inclusion of humans as a component of the landscape?

Landscape plays a very important role in these questions, necessitating the use of GIS data in our simulation. This article does not attempt to answer these questions; rather, these are the ultimate goals of the project that LiNK is designed to help answer.

A unique system of temples has existed on Bali for centuries; these temples and their associated forests act as refugia for the large populations of long-tailed macaques [12]. Each temple population consists of between 30 and 400 individuals. Existing behavioral and preliminary genetic evidence has documented the matrifocal society of the macaques, resulting in strong female philopatry [11], [12], [22]. Females remain at their birth temples and dominance is inherited maternally. Typically, subdominant and subadult males disperse from their natal temple (birth temple) around age seven, traveling to non-natal temple populations. Currently, actual dispersal distances and rates are unknown.

The ability of long-tailed macaques to coexist with humans has enabled a number of macaque populations to thrive in areas where other primate species have become extinct [12]. On Bali, human land-use patterns have resulted in a mosaic of riparian forest, small forest patches, agricultural lands, and urban areas across much of the island. The broad distribution of macaque populations on Bali suggests that the macaques

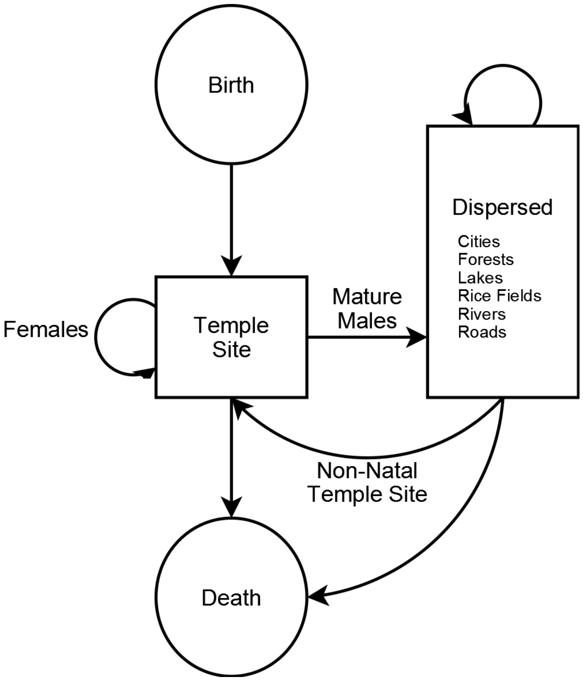


Fig. 5. Life cycle Transition Diagram. Macaques are always born in temple sites. Female macaques spend their entire lives within their natal temple. Mature male macaques disperse throughout the island through varying landscape with the ability to join other, non-natal, temples.

are utilizing the human modified landscape as it currently exists. Due to the protection and resource availability at temples, macaques are able to exist in moderately high densities alongside high density human populations. This co-existence, particularly surrounding the temples, has created an ideal study setting for evaluating how primate behavior and anthropogenic landscape changes influence pathogen transmission [10].

3.2. Conceptual Model

The conceptual model was developed by author Lane, with support from authors Fuentes and Hollocher. This group has closely studied macaques and an array of pathogens for a number of years. The basic model consists of a display of Bali with temple sites and macaques. We also display the contents of a given temple and provide the user with multiple model and pathogen parameter options. More detail on the components of the model follow.

Agents: Our agents are macaques, each with their own properties, such as location, sex, age, natal temple, and infection status. Macaques move in accordance to their surrounding environment, and males have the ability to enter and leave temples. Our model can support thousands of agents. We show a simplified transition diagram for the life cycle of our macaques in Figure 5.

Behavior: Macaques have the ability to move through their environment, interact with other macaques, reproduce, and die. Movement is dictated by their surrounding environment; macaques query their neighborhood and move appropriately. Macaques within a temple move randomly, with no GIS influence. All macaques have the ability to carry pathogens and can transmit pathogens when within a specified distance of one another. Reproduction is handled by allowing female macaques to produce offspring, with inherited traits, after they reach a specified age. As macaques age, they have a higher probability of dying.

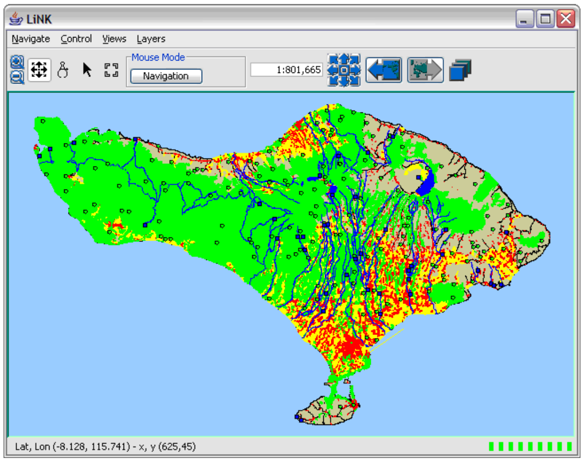


Fig. 6. LiNK Display. Here, we show Bali, Indonesia with all GIS landscape layers enabled, including the 42 temple sites. Macaques are shown as circles and temple sites as squares. Bali measures approximately 130km × 80km.

Interface: Researchers interact with the model through a simple control panel that allows them to tweak simulation parameters. Once the parameters are set, the user can begin running the simulation. The simulation is displayed via OpenMap, shown in Figure 6. Users can also see within temples.

Pathogens: LiNK has the ability to simulate a wide array of pathogens through the incorporation of several important pathogen parameters. The *infectivity* parameter refers to how infectious the pathogen being modeled is, while *virulence* is the proximity a macaque must be to another macaque to have the ability to transmit a pathogen. *Latency* represents how long a macaque takes to become symptomatic after becoming infected. *Acquired immunity* refers to the amount of time a macaque is immune to contracting a pathogen after having been previously infected. *Clearance time* is the amount of time a macaque takes to be cleared of a pathogen. Finally, *natural resistance* represents the proportion of macaques that are immune to a given pathogen. Selected pathogen-related variables and their temporal relationships are shown in Figure 7. A transition diagram for these variables is shown in Figure 8. Currently, LiNK has the ability to model one unique pathogen during a given simulation run.

Space: The macaques move about on 2D grids that represent temples sites and the island. The island grids are extrapolated from GIS data, at a customized granularity. For our purposes, a grid cell has sides of roughly 100m, leading to over one million possible locations. Each grid is called a layer; we have a total of eight layers: cities, forests, lakes, rice fields, rivers, roads, temples, and the actual island (called coast). These eight layers are melded together and use the same coordinate system. The coast and temple layers are required, while the rest can be turned on or off.

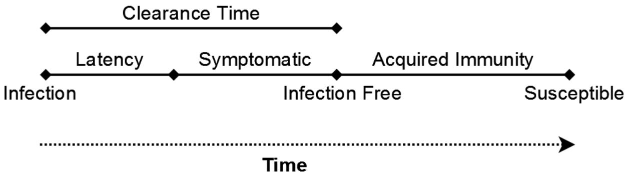


Fig. 7. Temporal Relationship of Pathogen Parameters and Related Events. The diagram above shows the relationship of the pathogen parameters in our simulation. Depending on the parameters used, macaques can become permanently immune to the modeled pathogen.

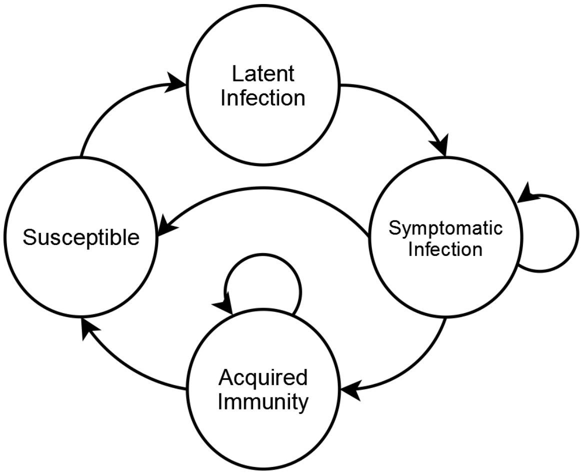


Fig. 8. Pathogen Transition Diagram. Macaques generally begin as *susceptible* and then transition to other states after being infected. Macaques with a symptomatic infection can become reinfected and macaques can reinfect themselves (autoinfection). An acquired immunity is gained after most infections, but may be lost after a given amount of time.

Time: One timestep in our simulation correlates to 12 real-world hours. Coupling this with 100m grid cells, we obtain the desired accuracy.

3.3. Implementation

There are several tools and technologies that made this study possible. The model is coded in Java with the Repast simulation toolkit [24]. We utilize Repast and OpenMap [23] to display the model and GeoTools [13] and JTS Topology Suite [17] to interact with the spatial information. The choice of tools used in this study was primarily driven by the necessity to process and visualize GIS data and to be cross-platform and open-source.

3.4. Verification and Validation

Simulations are useful only once they have passed some form of verification and validation. Verification refers to solving the model right, meaning that the simulation model matches the abstract model. Validation refers to solving the problem right, meaning the correct abstract model was chosen. ABMs must undergo and pass several subjective and quantitative verification and validation techniques to be considered valuable models [2], [3], [21], [28]. Figure 9 shows common techniques for ABMs, adapted from Kennedy, et al. [19].

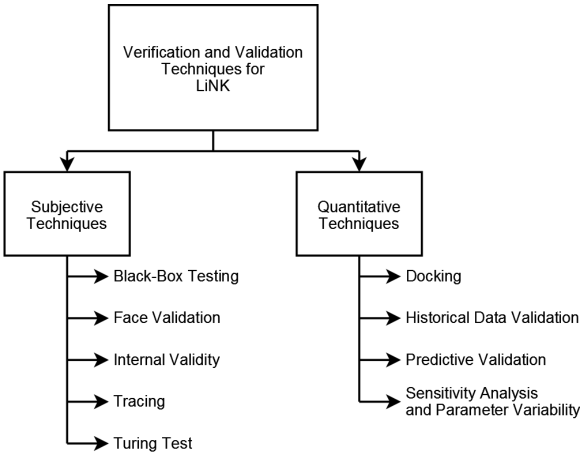


Fig. 9. Verification and Validation Techniques. Here, we show techniques we used and plan to use for the verification and validation of LiNK.

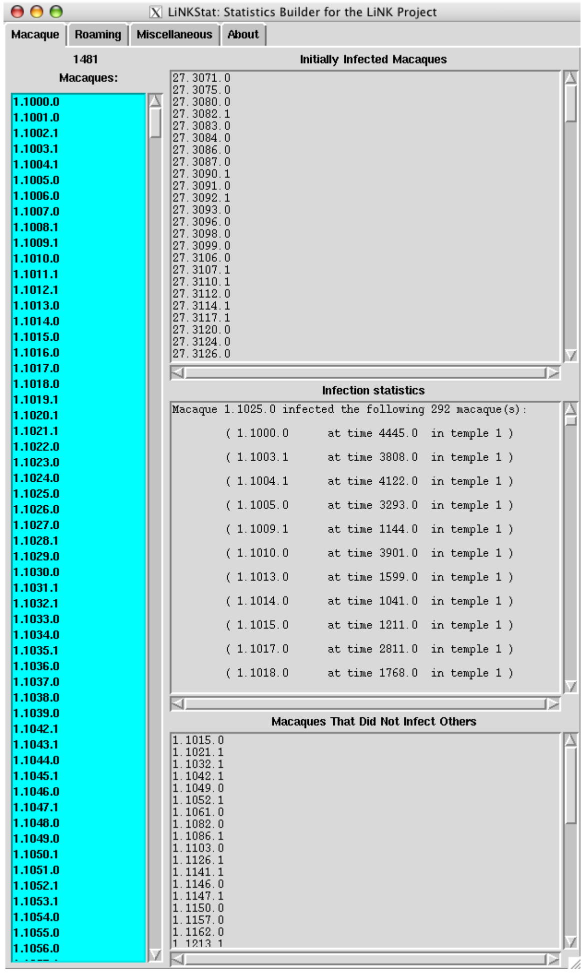
The LiNK model was developed in conjunction with domain experts from multiple fields and has undergone extensive face validation, both through its display and evaluation of its output. We have also checked for internal validity and traced entities of the model. Much of this work has been performed through the use of LiNKStat, which we next describe. We are currently collecting more real-world data that we will use in conjunction with the current data to continue docking LiNK. LiNK's predictive power has also been considered, and we are planning real-world experiments to evaluate this.

3.5. LiNKStat

LiNK is a complex model; as such, it creates enormous amounts of output. To glean scientific insight and validation, LiNK tracks of a wide array of events, including infections, births, deaths, and when a macaque enters or leaves a temple. When simulations are run over a long period of time, it is not uncommon to have tens of millions of events, or more. We have created an interactive graphical tool, LiNKStat, to analyze output from LiNK. LiNKStat parses through output files and builds graphs to gather statistics about the model. For example, LiNKStat allows users to track the route of infection from a given macaque, obtaining statistics such as number of macaques directly or indirectly infected. Such statistics help subject matter experts collect insight from LiNK. A screen capture of LiNKStat is shown in Figure 10 and an example graph from its output is shown in Figure 11. LiNKStat is efficient, with a runtime mainly dependent on the number of infection events and their degree of proliferation.

3.6. Performance

The model has utilized the aforementioned techniques to interact with GIS data. We started with hefty raster-based queries and refined our method until we achieved the balance of specificity and speed we desired. Table I shows the initial GUI load time for the model for each technique, and Table II and Figure 12 show the number of timesteps simulated per second for each query mechanism. These tables and figure



1481

Macaques:

```
1.1000.0
1.1001.0
1.1002.1
1.1003.1
1.1004.1
1.1005.0
1.1006.0
1.1007.0
1.1008.1
1.1009.1
1.1010.0
1.1011.1
1.1012.1
1.1013.0
1.1014.0
1.1015.0
1.1016.0
1.1017.0
1.1018.0
1.1019.1
1.1020.1
1.1021.1
1.1022.0
1.1023.0
1.1024.0
1.1025.0
1.1026.0
1.1027.0
1.1028.1
1.1029.0
1.1030.0
1.1031.1
1.1032.1
1.1033.0
1.1034.0
1.1035.1
1.1036.0
1.1037.0
1.1038.0
1.1039.0
1.1042.1
1.1043.1
1.1044.0
1.1045.1
1.1046.0
1.1047.1
1.1048.0
1.1049.0
1.1050.1
1.1051.0
1.1052.1
1.1053.1
1.1054.0
1.1055.0
1.1056.0
```

Initially Infected Macaques

```
27.3071.0
27.3075.0
27.3080.0
27.3082.1
27.3083.0
27.3084.0
27.3086.0
27.3087.0
27.3090.1
27.3091.0
27.3092.1
27.3093.0
27.3096.0
27.3098.0
27.3099.0
27.3106.0
27.3107.1
27.3110.1
27.3112.0
27.3114.1
27.3117.1
27.3120.0
27.3124.0
27.3126.0
```

Infection statistics

```
Macaque 1.1025.0 infected the following 292 macaque(s):

        ( 1.1000.0      at time 4445.0  in temple 1 )

        ( 1.1003.1      at time 3808.0  in temple 1 )

        ( 1.1004.1      at time 4122.0  in temple 1 )

        ( 1.1005.0      at time 3293.0  in temple 1 )

        ( 1.1009.1      at time 1144.0  in temple 1 )

        ( 1.1010.0      at time 3901.0  in temple 1 )

        ( 1.1013.0      at time 1599.0  in temple 1 )

        ( 1.1014.0      at time 1041.0  in temple 1 )

        ( 1.1015.0      at time 1211.0  in temple 1 )

        ( 1.1017.0      at time 2811.0  in temple 1 )

        ( 1.1018.0      at time 1768.0  in temple 1 )
```

Macaques That Did Not Infect Others

```
1.1015.0
1.1021.1
1.1032.1
1.1042.1
1.1049.0
1.1052.1
1.1061.0
1.1082.0
1.1086.1
1.1103.0
1.1126.1
1.1141.1
1.1146.0
1.1147.1
1.1150.0
1.1157.0
1.1162.0
1.1213.1
```

Fig. 10. LiNKStat. This screen capture shows one of the analysis tabs of LiNKStat. The left column displays an interactive list of macaques in the simulation that updates the middle right panel with specific infection statistics. These statistics form graphs, an example of which is shown in Figure 11. LiNKStat has been and will continue to be very helpful in the verification and validation of LiNK.

TABLE I
PERFORMANCE COMPARISON

	GUI Load Time (s)	
	Coast, Lakes	Coast, Lakes, Forests
Spatial Query	3.5	3.5
Raster Query	35	42
Simplified Spatial Query	1.8	2.5
Precalculated Query Matrix	1.6	2

TABLE II
PERFORMANCE COMPARISON

	Timesteps/sec	
	Coast, Lakes	Coast, Lakes, Forests
Spatial Query	1.6	0.15
Raster Query	18.5 (11x faster)	19 (126x)
Simplified Spatial Query	39.5 (25x)	15.8 (105x)
Precalculated Query Matrix	126.2 (79x)	124.1 (827x)
Precalculated Query Matrix, non-GUI	669.6 (419x)	650.2 (4335x)

show averages with either the coast and lakes or the coast, lakes, and forests layers enabled, all with the same number of initial agents. Spatial queries were predictably slowest, as the raw vector files contain an enormous amount of realism, making calculations expensive. Utilizing raster data offers a significant improvement but with the drawback of the long initial startup time. Our simplified spatial query greatly improves upon the traditional spatial query, but performance drops significantly as more layers are added. Utilizing precalculated query matrices produces the fastest simulation, with even greater gains when the display is disabled. Table III and its corresponding Figure 13 show the scalability, in terms of number of agents, for the raster and precalculated query matrix method. The precalculated query matrix method scales very well as the amount of GIS data increases and adequately as the number of agents increases. The precalculated query matrix method offers the best, scalable results. All performance tests were run on a single core as a single thread on a Core 2 Duo 2.0 GHz laptop, highlighting further potential in scalability. Numbers listed in the figures are averages of 10 simulation runs. Additionally, LiNK has been ported to run on a high-performance computing cluster, making it easy to automate, greatly increasing its utility.

3.7. *Results*

LiNK has demonstrated the importance of landscape in the scope of epidemiological modeling [22]. The model has been improved in terms of speed and scalability through an abstraction of typical GIS data representation. We have shown the ability to have many agents interact with complex spatial data in a time frame adequate for a simulation. Additionally, we have begun to show the impact of landscape on pathogen transmission, which is thus far in accordance with real-world data from Roberts and Janovy [25]. Further sensitivity analysis and more verification and validation needs to be performed.

4. DISCUSSION

When designing an ABM with GIS aware agents, there are a number of factors that should be considered. Scalability in terms of the number of agents is probably the most important factor to consider. Other important issues include the complexity of the GIS data and the amount of GIS data that the model will rely upon. An adept modeler will utilize the GIS data at a granularity appropriate for the model at hand. In

TABLE III
SCALABILITY COMPARISON

	Timesteps/s		
Number of Initial Dispersed Macaques	10	100	1000
Raster Query (3 Layers)	51.3	29	19.9
Raster Query (7 Layers)	33.6	27.6	11
Precalculated Query Matrix (3 Layers)	140.7	131.4	83.8
Precalculated Query Matrix (3 Layers)	137.5	129.5	82.9
Precalculated Query Matrix, non-GUI (3 Layers)	669.8	487.8	154
Precalculated Query Matrix, non-GUI (7 Layers)	680.4	529.6	158.2

terms of speed, raster data scales reasonably with increasing GIS complexity, but not as well with an increase in the number of agents. Spatial queries scale poorly with an increase in the amount of GIS data and complexity, as well as with an increase in the number of agents. Regarding accuracy, utilizing vector data via spatial queries offers the highest accuracy, but at the highest performance cost. Raster data and our precalculated query matrix method offer varying levels of accuracy, while offering faster speed. Table IV summarizes general ratings for each approach. Possible ratings are 1-5, from *Poor* to *Excellent*. Accuracy of GIS data refers to the faithfulness to the original GIS data, while the amount of GIS data refers to the ability of each technique to handle multiple layers of GIS data. The remaining metrics are self-explanatory. Our precalculated query matrix method scales best in terms of number of agents and particularly in the amount of GIS data present.

5. Conclusion

We have presented a complex model of pathogen transmission that utilizes GIS data. This model has begun to demonstrate the importance of integrating spatial data into models of pathogen transmission. We have created an efficient and effective mechanism to allow our agents to become GIS aware. Future extensions to the model include adding the ability to model different pathogens simultaneously, deploying a web-based front end to the model, and allowing for the use of custom GIS data. We would also like to explore running our simulation on graphics processing units, as described in D'Souza, et al. [9]. Finally, we plan to further verify and validate the LiNK model through real-world data.

Acknowledgment

The authors would like to thank the reviewers and organizers for this special issue. We would like to thank Gerhard Niederweiser for his work in helping design the precursor to the LiNK model. We would also like to thank the National Science Foundation and the University of Notre Dame's Institute for Scholarship in the Liberal Arts for funding on this project.

References

[1] S. M. Anwar, M. Musiani, G. McDermid, and D. Marceau. How Do Human Activities Shape Wolves' Behavior In The Central Rocky Mountain Region, Alberta, Canada? In L. Yilmaz, editor, *Proceedings of the 2009 Agent-Directed Simulation Symposium*. The Society for Modeling and Simulation International, March 2009.

[2] O. Balci. *Handbook of Simulation: Principles, Methodology, Advances, Applications, and Practice*, chapter Verification, Validation, and Testing. John Wiley & Sons, New York, NY, 1998.

[3] J. Banks and J. S. C. II. Introduction to discrete-event simulation. In *Proceedings of the 1986 Winter Simulation Conference*, pages 17–23, 1986.

[4] D. Brown, R. Riolo, D. Robinson, M. North, and W. Rand. Spatial Process and Data Models: Toward Integration of Agent-based Models and GIS. *Journal of Geographic Systems, Special Issue on Space-Time Information Systems*, 7(1):25–47, 2005.

[5] C. Castle, A. Crooks, P. Longley, and M. Batty. Agent-Based Modelling and Simulation using Repast: A Gallery of GIS Applications from CASA. In G. Priestnall and P. Alpin, editors, *Proceedings of the 14th Geographical Information Systems Research UK Conference*, pages 237–239, 2006.

[6] A. T. Crooks. UCL Working Paper Series: The Repast Simulation/Modelling System for Geospatial Simulation. available at http://www.casa.ucl.ac.uk/working_papers/paper123.pdf, September 2007.

[7] P. Daszak, A. Cunningham, and A. Hyatt. Anthropogenic environmental change and the emergence of infectious disease in wildlife. *Acta Tropica*, 78:103–116, 2001.

[8] Douglas-Peucker Algorithm. http://geometryalgorithms.com/Archive/algorithm_0205/#Douglas-Peucker%20algorithm.

[9] R. M. D'Souza, M. Lysenko, S. Marino, and D. Kirschner. Data-Parallel Algorithms for Agent-Based Model Simulation of Tuberculosis On Graphics Processing Units. In L. Yilmaz, editor, *Proceedings of the 2009 Agent-Directed Simulation Symposium*. The Society for Modeling and Simulation International, March 2009.

[10] L. J. Engel, G. A. Engel, M. A. Schillaci, A. Rompis, A. Putra, K. G. Suaryana, A. Fuentes, B. Beer, S. Hicks, R. White, B. Wilson, and J. S. Allan. Primate-to-Human Retroviral Transmission in Asia. *Emerging Infectious Diseases*, 11(7), July 2005.

[11] J. E. Fa and D. G. Lindburg, editors. *Evolution and Ecology of Macaque Socities*. Cambridge University Press, 2005.

[12] A. Fuentes, M. Southern, and K. G. Suaryana. Monkey forests and human landscapes: Is extensive sympatry sustainable for *Homo sapiens* and *Macaca fascicularis* on Bali? In J. D. Patterson and J. Wallis, editors, *Commensalism and Conflict: The Primate-Human Interface*. American Society of Primatology Publications, 2005.

[13] GeoTools. http://geotools.codehaus.org.

[14] N. Gilbert. *Agent-based Models*. SAGE Publications, Thousand Oaks, CA, 2008.

[15] H. R. Gimblett. Integrating geographic information systems and agent-based technologies for modeling and simulating social and ecological phenomena. In H. R. Gimblett, editor, *Integrating Geographic Information Systems and Agent-based Modeling Techniques for Simulating Social and Ecological Processes*. Oxford University Press, 2002.

[16] V. Grimm and S. F. Railsback. *Individual-based Modeling and Ecology*. Princeton University Press, Princeton, NJ, 2005.

[17] JTS Topology Suite. http://www.vividsolutions.com/jts/jtshome.htm.

[18] M. Keeling, M. Woolhouse, R. May, G. Davies, and B. Grenfell. Modelling vaccination strategies against foot-and-mouth disease. *Nature*, 421:136–142, January 2003.

[19] R. C. Kennedy. Verification and Validation of Agent-based and Equation-based Simulations and Bioinformatics Computing: Identifying Transposable Elements in the *Aedes aegypti* Genome. Master's thesis, University of Notre Dame, April 2006.

[20] R. C. Kennedy, K. E. Lane, A. Fuentes, H. Hollocher, and G. Madey. Spatially Aware Agents: An effective and efficient use of GIS data within an Agent-based Model. In L. Yilmaz, editor, *Proceedings of the 2009 Agent-Directed Simulation Symposium*. The Society for Modeling and Simulation International, March 2009.

[21] R. C. Kennedy, X. Xiang, T. F. Cosimano, L. A. Arthurs, P. A. Maurice, and S. E. Cabaniss. Verification and Validation of Agent-based and Equation-based Simulations: A Comparison. In L. Yilmaz, editor, *Proceedings of the 2006 Agent-Directed Simulation Symposium*. The Society for Modeling and Simulation International, April 2006.

[22] K. E. Lane, R. C. Kennedy, L. A. Miller, G. Madey, H. Hollocher, and A. Fuentes. Exploring the use of agent-based models in understanding patterns of pathogen transmission. *American Journal of Primatology*, 2009. In prep.

[23] OpenMap. http://openmap.bbn.com.

[24] Repast. http://sourceforge.repast.net.

[25] L. Roberts and John Janovy, Jr. *Gerald D. Schmidt & Larry S. Roberts' Foundations of Parasitology*. McGraw-Hill, eighth edition, 2009.

[26] S. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Pearson Education, Inc., Upper Saddle River, NJ, 2003.

[27] L. Temime, Y. Pannet, L. Kardas, L. Opatowski, D. Guillemot, and P. Y. Boëlle. NOSOSIM: an agent-based model of pathogen circulation in a hospital ward. In L. Yilmaz, editor, *Proceedings of the 2009 Agent-Directed Simulation Symposium*. The Society for Modeling and Simulation International, March 2009.

[28] X. Xiang, R. Kennedy, G. Madey, and S. Cabaniss. Verification and Validation of Agent-based Scientific Simulation Models. In L. Yilmaz, editor, *Proceedings of the 2005 Agent-Directed Simulation Symposium*, volume 37, pages 47–55. The Society for Modeling and Simulation International, April 2005.

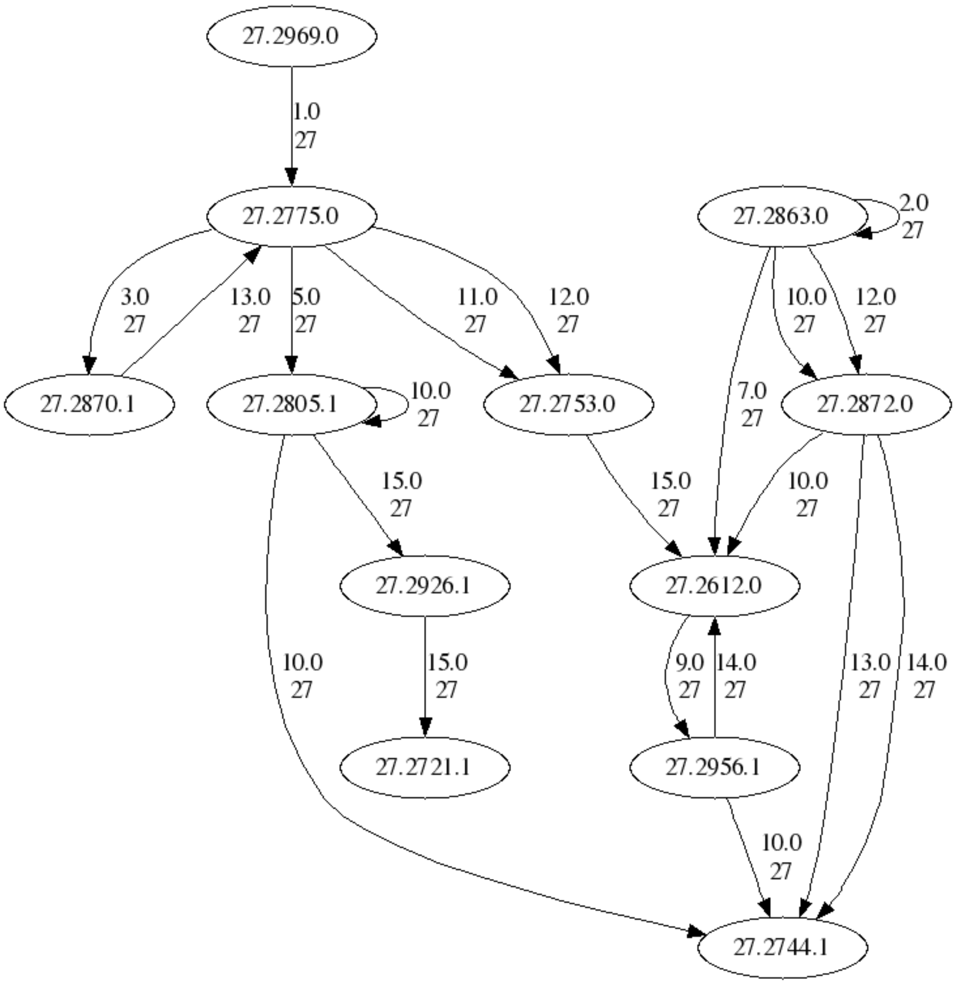


Fig. 11. LiNKStat Pathogen Transmission Graph. The graph above allows us to visually track pathogen transmission, helping with validation and interpretation of output. Nodes refer to macaques, with the naming convention being natal temple number concatenated with an id concatenated with a sex identifier. For example, the topmost node would be parsed as a female macaque with temple 27 as its natal temple and 2969 as its id. Transitions are infection events, listed with the timestep and location where the infection occurred. Starting at the top, macaque 27.2969.0, infected macaque 27.2775.0 at timestep 1, in temple 27. Macaque 27.2775.0 went on to infect four other macaques, and was also reinfected by macaque 27.2870.1. Autoinfection is possible as indicated by nodes 27.2863.0 and 27.2805.1.

TABLE IV
ADVANTAGES AND DISADVANTAGES (1- POOR; 5- EXCELLENT)

	Raster Query	Spatial Query	Simplified Spatial Query	Precalculated Query Matrix
Accuracy of GIS Data	3	5	4	4
Amount of GIS Data	3	1	2	5
Complexity of GIS Data	2	5	4	4
Load Time	1	4	4	5
Memory Requirement	2	4	4	5
Number of Agents	4	1	2	4
Timesteps/s	4	1	2	5

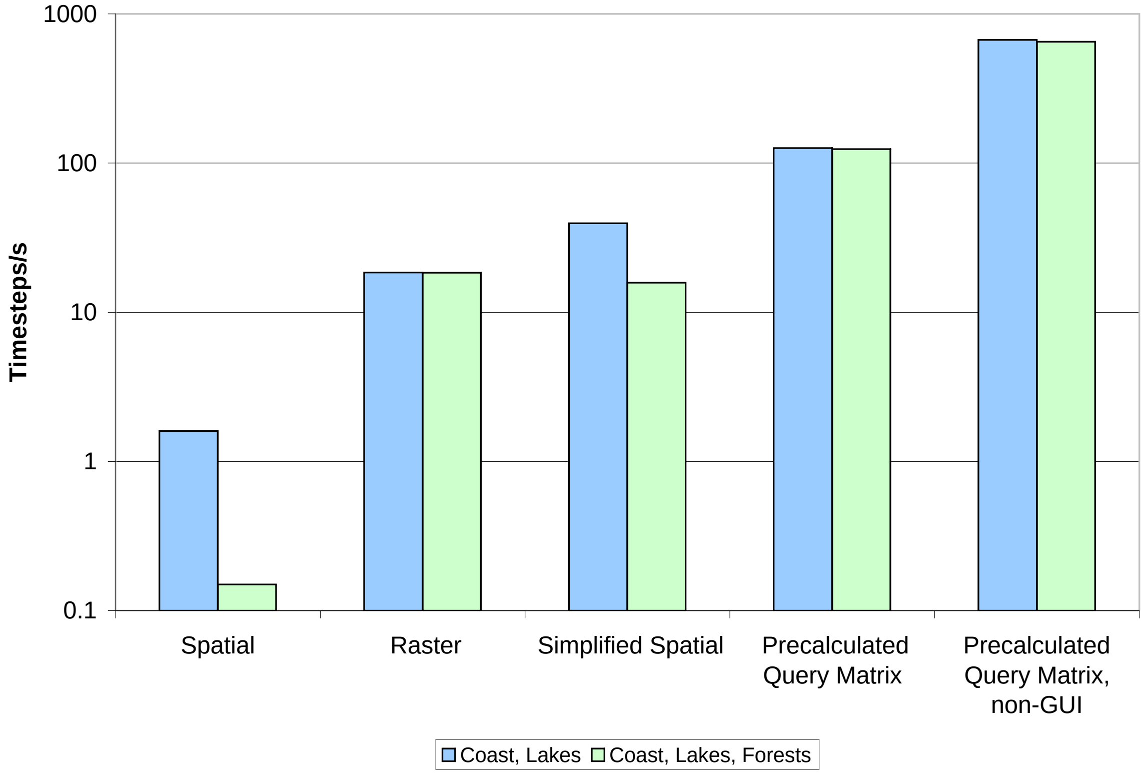


Fig. 12. Performance comparison of varying query methods. The figure shows that we obtained nearly an order of magnitude in terms of speed in going from spatial to raster to simplified spatial queries, and then almost another order of magnitude from raster to simplified spatial queries. Finally, disabling the GUI offers nearly another order of magnitude improvement. It is also notable that enabling more layers in non-GUI mode adds almost no performance hit. We show the figure above with a logarithmic scale.

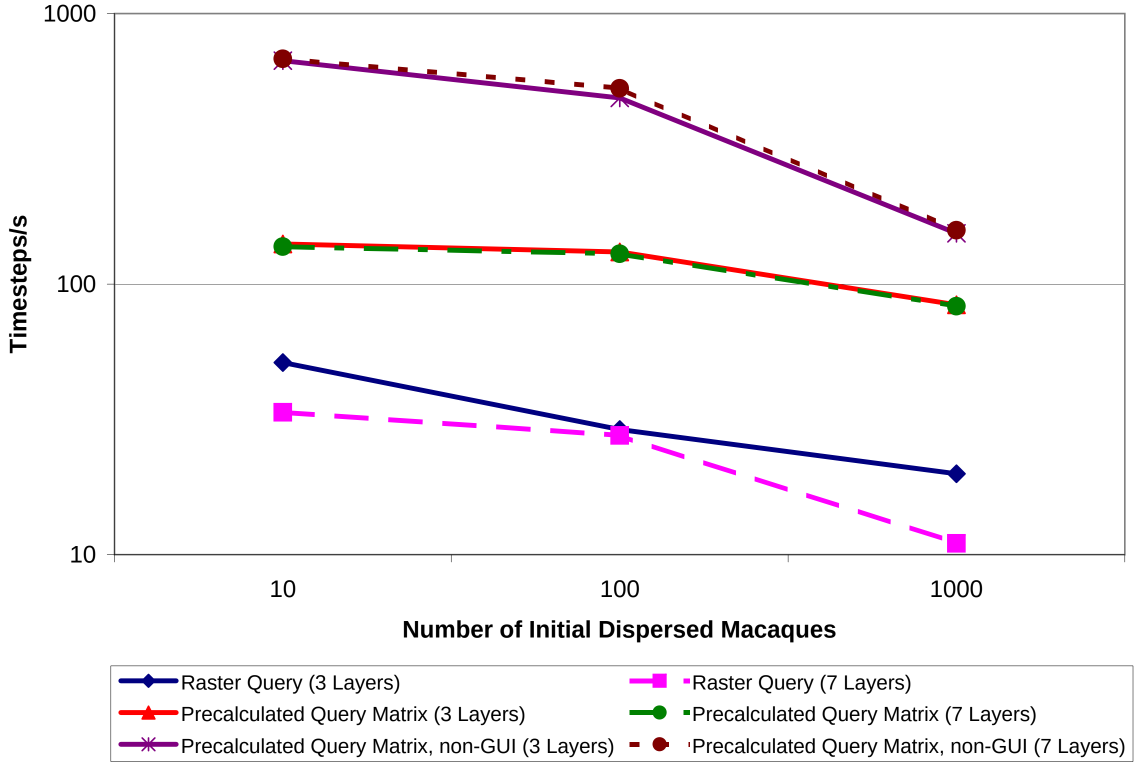


Fig. 13. Scalability with respect to initial number of dispersed macaques and amount of GIS data. Here, we show simulations starting with 10, 100, and 1000 dispersed macaques across different querying mechanisms. The precalculated query matrix method performs best in all cases, even better with 1000 agents than other methods with 10 agents. The figure is shown on a logarithmic scale.

Ryan C. Kennedy received his B.S. and M.S degrees from the University of Notre Dame in 2004 and 2006, respectively. He is currently working on his Ph.D. in Computer Science and Engineering, also at the University of Notre Dame. His research interests include agent-based simulations, bioinformatics, with a focus on computational pipelines for the discovery and annotation of transposable elements, and verification and validation of agent-based simulations. He is a member of ACM, IEEE, and the Society for Computer Simulation.

Kelly E. Lane received her B.S. from the University of Denver, her M.S. from Saint Louis University, and is currently a Ph.D. candidate at the University of Notre Dame. Her research focuses on host-parasite dynamics and the ecology and evolution of wildlife infectious diseases. In past research, she has examined the effects of population size and developmental stress on the morphology and parasite burden of *Vespertilionidae* bats. Currently, she is exploring the ecological drivers of pathogen emergence by investigating the influence of anthropogenic landscapes on gastrointestinal parasite dynamics and population structure of primates in Southeast Asia.

S. M. Niaz Arifin received his B.S. from Bangladesh University of Engineering and Technology (BUET) in 2004 and his M.S. from the University of Texas at Dallas in 2006. He is currently working on his Ph.D. at the University of Notre Dame. His research interests include agent-based simulations, mathematical modeling in biology and data mining.

His M.S. research focus was on artificial intelligence and natural language processing. He served as a software developer in the stereotactic breast cancer treatment project at Xcision Medical Systems, California and the Rails online database project at Sabre Holdings Corporation, Texas.

Agustín Fuentes completed a B.A. in Zoology and Anthropology, and an M.A. & Ph.D. in Anthropology at the University of California, Berkeley. He taught in the Department of Anthropology and directed the Primate Behavior and Ecology Program at Central Washington University from 1996-2002 and is currently Professor of Anthropology at the University of Notre Dame. His research and teaching interests include the evolution of social complexity in human and primate societies, cooperation and conflict negotiation across primates, including humans, and reproductive behavior and ecology. He is also interested in issues of human-nonhuman primate interactions, disease and pathogen transfer. Fuentes' recent published work includes the books "Evolution of Human Behavior" (Oxford University press) "Core Concepts in Biological Anthropology" (McGraw-Hill) and "Primates in Perspective" (co-edited, Oxford University Press) and articles such as "It's Not All Sex and Violence: Integrated Anthropology and the Role of Cooperation and Social Complexity in Human Evolution" and "The humanity of animals and the animality of humans: A view from biological anthropology inspired by J.M. Coetzees' Elizabeth Costello" in the *American Anthropologist*, and "Human culture and monkey behavior: Assessing the contexts of potential pathogen transmission between macaques and humans" in the *American Journal of Primatology*. His current research projects include assessing behavior, ecology, and pathogen transmission in human-monkey interactions in Southeast Asia and Gibraltar and examining the roles of cooperation, social negotiation, and patterns of niche construction in primate and human evolution.

Hope Hollocher received a B.A. in Biology from the University of Pennsylvania and a Ph.D. in Population and Evolutionary Genetics from Washington University in St. Louis. After working as a postdoctoral fellow at the University of Chicago, she taught in the Department of Ecology, Evolution and Behavior at Princeton University from 1994-2000 and is currently an Associate Professor in the Department of Biological Sciences at the University of Notre Dame. Her main research focus is on genetic mechanisms underlying speciation, evolution and development, and landscape genetics, and she has published extensively on those topics in a variety of journals including, *Evolution, Genetics, Molecular Biology and Evolution, Journal of Experimental Zoology, Genetical Research, Heredity, Nature, and the Proceedings of the National Academy of Sciences*. She has received recognition for her work as the recipient of the Alfred P. Sloan Young Investigators Award in Molecular Studies of Evolution and held the Clare Boothe Luce Collegiate Chair in Biology at the University of Notre Dame from 2000-2005. She is an active member of several scientific societies and has held appointed and elected positions in the Society for the Study of Evolution, the Society for Molecular Biology and Evolution, and the American Association for the Advancement of Science. Her most recent research incorporates population genetic theory into issues of disease ecology and investigates how genetic structuring of host populations influences the transmission and differentiation of pathogens in primates.

Gregory R. Madey received the Ph.D and M.S. degrees in operations research from Case Western Reserve University and the M.S. and B.S degrees in mathematics from Cleveland State University. He worked in industry for several firms, including Goodyear Aerospace, Gould Oceans Systems (now part of Northrup-Grumman), and Loral (now part of Lockheed Martin). He is currently faculty in the Department of Computer Science and Engineering at the University of Notre Dame. His research includes topics in agent-based modeling and simulation, emergency management modeling and simulation, web-services and service oriented architectures, bioinformatics, web portals for scientific collaboration, open source software, and cyberinfrastructure. He has published in various journals including, *Communications of the ACM, IEEE Transactions on Engineering Management, IEEE Computing in Science & Engineering, The Journal of Systems & Software, BMC Bioinformatics, Computational & Mathematical Organization Theory, Nucleic Acids Research, Decision Sciences, The European Journal of OR, Omega, Expert Systems with Applications, and Expert Systems*. He is a member of the ACM, AIS, IEEE Computer Society, Informs, and the Society for Computer Simulation.