Modeling Language Diffusion With ArcGIS

Prepared for WORLDMAP.ORG
JUNE 2004

Mr. Christopher Deckert
Campus Crusade for Christ
The JESUS Film Project
PO BOX 72007
San Clemente, CA 92674-2007
Tel.: 949-361-7575

Table of Contents

Abstract

Currently there are more than six thousand eight hundred languages across the globe. More than four hundred of those languages are nearly extinct, and many more are on the endangered list. As we move into a more global economy, interactions between cultures occur daily. Businesses, governments, and nongovernmental organizations bring goods and services to every facet of society in every part of the earth. Knowing the language of the intended recipient is essential for smooth transactions. Taking the current sources of language information, we have developed a predictive model for language diffusion. Using ArcGIS, ArcGIS Spatial Analyst, ArcObjects, and Model Builder the relative cost of language propagation was generated. Anthropogenic factors produced an ethnic cost grid. Various environmental and geographic factors produced a physical cost grid. These grids were combined to approximate the probability of any language spoken in any given location on the globe.

Major Data Sources

[1]Languages of the world shapefile with approximately 7,730 polygons from the Language Mapping Project of Global Mapping International with SIL.

[2]Oak Ridge National Laboratory LandScan Global Population 2001 Database—Resolution 1 km grid

[3]GTOPO digital elevation model (DEM) for the entire world in grid format—resolution 1 km (926 m)

[4] DCW Road coverage of the world with at least two categories of roads – 1:1,000,000

[5]DCW River coverage of the world – 1:1,000,000

[6]DCW Railroad coverage of the world – 1:1,000,000

[7]DCW Countries coverage of the world – 1:1,000,000

[8]National Imagery and Mapping Agency (NIMA—now National Geospatial–Intelligence Agency [NGA]) populated places point shapefile – various resolutions

[9]Ethnologue Language Family Tree data – Ethnologue.com

Introduction

The modern geographic information system (GIS) such as the ArcGIS, offers extensive functionality and allows for full customization of a wide variety of applications. Customization options enable GIS professionals to generate intuitive and easy-to-use interfaces. Behind an attractive user interface, complex algorithms coded using industry-standard Component Object Model (COM)/ArcObjects are hidden. Such an approach was utilized for this model to determine the "diffusion" of a language (an approximated probability of a particular language being spoken outside of its current area as indicated in the Language Mapping Project[1]), utilizing numerous functions offered by ArcGIS. With this, a surface representing the "cost" of language propagation was generated. For this model, such a cost surface was derived from two major sources. The identified constituents included: (1) an "ethnic" cost surface based on anthropogenic parameters, and (2) a "physical" cost surface derived from physical geography criteria. Each one of these two major surfaces was constructed based on several relevant components.

The concepts included in the developed algorithm were tested using data for the countries of Thailand and Nigeria. To run the individual country approach, such as performed for this model all the component grids were clipped to the extent of the polygon(s) of a particular language and buffered by 200-km. Actually, the size of the buffer should be proportional to the language's "altitude" or "weight" (a combination of the number of speakers, fertility rate, average education level, average income level of the language speakers, etc.). However, because of the lack of the mentioned data, for this model we applied a standard 200 km wide buffer. For many languages, all the input data should be buffered around each country's borders, so the data extends beyond the international borders of every individual country. In other words, the zone of a particular language influence should not be limited to the political borders of a country.

1.0 ETHNIC Cost Surface Grid

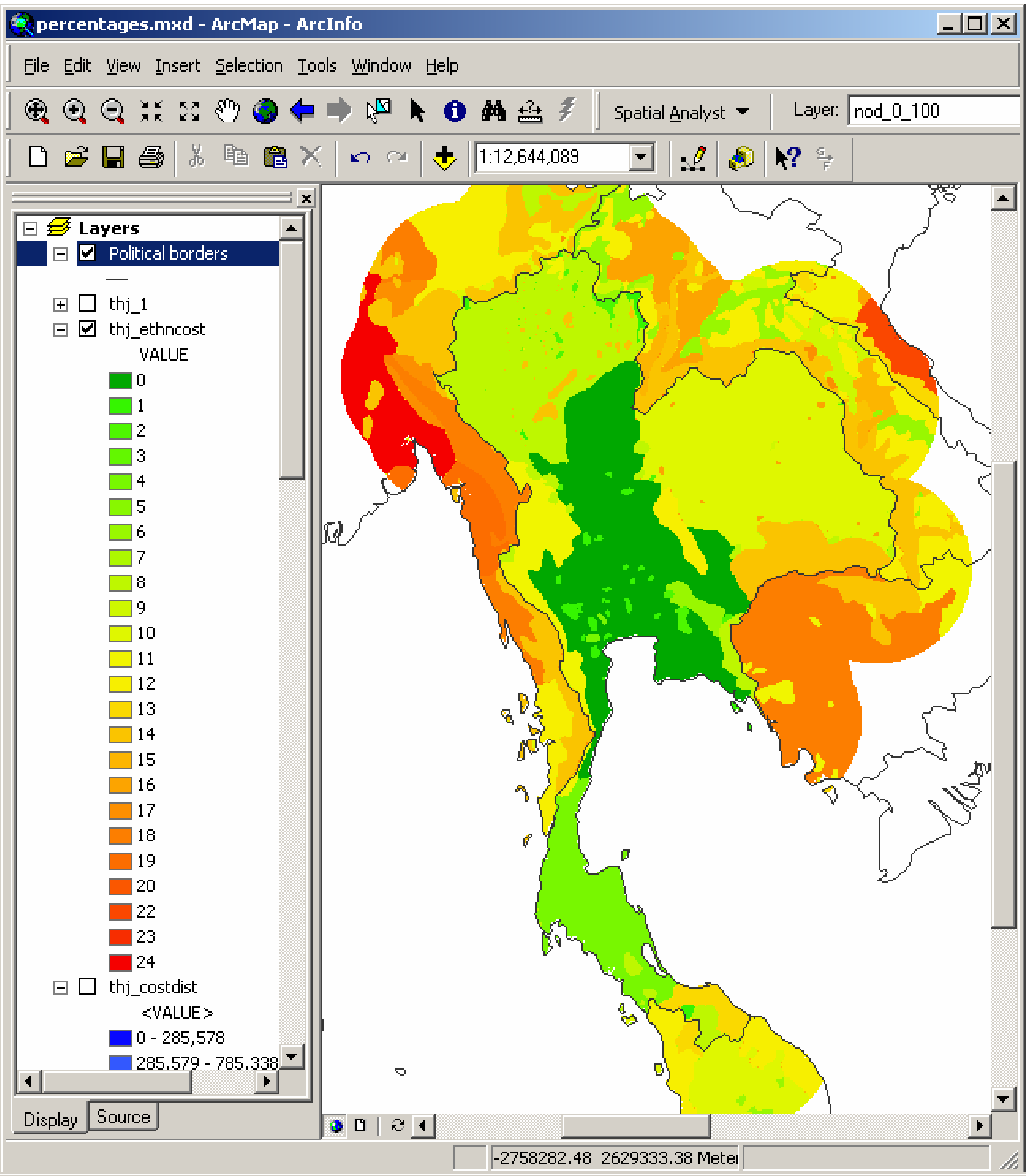
The foundations of the ETHNIC cost surface are the grids representing language popularity, the linguistic proximity. The following paragraphs describe the content of the preliminary grids (A through C) used to compose the final ETHNIC cost grid.

A. The language popularity grid is based on the number of speakers in every linguistic group, as represented in the original shapefile of language polygons[1]. This grid could be generated just once for each country by intersecting the polygons defining the extent of languages[1] with the Landscan Global Population data[2]. The resulting linguistic popularity grid was then reclassified into 10 categories using the SLICE function of ESRI's ArcGIS Spatial Analyst extension to ArcMap. Finally, the considered language was attributed with the lowest value (least cost) of 0 on the final output language popularity grid.

B. The linguistic proximity grid is based on the Language Family Trees[9] (LFT), as published by the Ethnologue organization (www.ethnologue.com). These grids are created individually for every language that is an input to the model. The values (representing the "linguistic distances" between the reviewed language and all other languages spoken within the above-mentioned buffer of 200 km) need to be of the integer type and will range from 0 (zero reserved for the language currently reviewed) to 10. The maximum linguistic distance between any two of the most remote languages on the LFT is presently 7. However, the LFT does not provide classification beyond the top 20 fundamental linguistic family groups. These are also hierarchically classified by other sources into the additional three levels. The typical maximum value of the linguistic proximity grid will be between 3 and 5, depending on the level of ethnic diversity in a given country. The examples of countries where the proximity values reach extremes are India and Russia with the highest values (8–10), and Japan and Korea with the lowest values (1–2).

C. The combined final "ethnic" cost surface grid is a result of a simple addition using the grid algebra of grids A and B described above. The value for the currently studied language would be 0, while the highest theoretically possible values might reach a value of 20. Most likely, the maximum values will not exceed 15. The values of this ethnic grid are proportional to the "cost" of the language propagation derived from the ethnopolitical criteria. These ethnic cost surface grids will be unique for every language.

Figure 1-1


percentages.mxd - ArcMap - ArcInfo
File Edit View Insert Selection Tools Window Help
Spatial Analyst
Display
Source

Layer: nod_0_100

1:12,644,089

- Layers
  - ☑ Political borders
  - □ thj_1
  - ☑ thj_ethncost
    - VALUE
    - 🟩 0
    - 🟩 1
    - 🟩 2
    - 🟩 3
    - 🟩 4
    - 🟩 5
    - 🟩 6
    - 🟩 7
    - 🟩 8
    - 🟨 9
    - 🟨 10
    - 🟨 11
    - 🟨 12
    - 🟨 13
    - 🟧 14
    - 🟧 15
    - 🟧 16
    - 🟧 17
    - 🟧 18
    - 🟧 19
    - 🟥 20
    - 🟥 22
    - 🟥 23
    - 🟥 24
  - □ thj_costdist
    - <VALUE>
    - 🟦 0 - 285,578
    - 🟦 285,579 - 785,338

-2758282.48 2629333.38 Meter

The aggregated ethnic cost surface for the Thai Central Language (dark green).

2.0 PHYSICAL Cost Surface Grid

The PHYSICAL cost surface grid is a weighted merge of the five grids (D,E,F,G, and H) described below (D–G). Relative weights may be assigned by a user to each one of the grids to reflect their significance in local geographic and socioeconomic conditions.

D. Proximity to roads—The proximity to roads grid C was generated by buffering roads by 40 km for the main roads and 20 km for the secondary or smaller roads. To accomplish this, the ArcGIS Spatial Analyst function EUCDISTANCE was run to create the values of the grid. Assuming that the cost of moving over the small roads would be about twice as high as the major roads, the result of the EUCDISTANCE for the small roads was multiplied by 2 before merging it with the output of the EUCDISTANCE for the main proximity to roads cost surface.

E. Proximity to rivers (and lakes)—Proximity to rivers (and lakes) was created similarly to the roads approach. Since it is much harder (but possible) to classify rivers, we used just one category for rivers. Steps necessary to generate the proximity to rivers grid include the following:

1. Rasterize the rivers with the resolution of 1 km, with the cells representing rivers having the value of 1, and all the other cells classified as NoData cells.

2. Use the output of the step 1 as the <source> grid to the EUCDISTANCE function.

3. Specify the {max-distance} (e.g., 30,000 meters). This value should be considered a variable to be set by a user specifically for each region of the world, depending on the significance of the river network in terms of transportation and communication.
```
Rivercost = eucdistance ( rivers, #, #, 30000 )
```

F. Proximity to railroads—Proximity to the railroads grid was derived from the existing Digital Chart of the World originally designed and completed by ESRI at the scale of 1:1 million. The data contain a global network of railroad lines.

For the existing railroads, a diminishing buffer of 50 kilometers was generated by running the EUCDISTANCE function to represent the zone of railway line influence and the role in population migration and communication.

G. Geographic latitude and elevation above sea level, substituting for the climate conditions—Elevation larger than a certain value (depending on the climatic conditions) acts as a barrier for people migrating, and consequently, it constitutes impedance for language propagation. To implement a representation of this kind of a barrier, a modified elevation grid was generated to reduce the impedance of elevations at low latitudes and to increase it at high latitudes. The steps below describe how this was calculated.

1. The trigonometric function of SIN available in the ArcGIS Spatial Analyst extension was run on the grid representing the latitudes. This function creates a grid with values representing relative angles of sun radiation. More precisely, the function

assigns relative values to latitudes ranging from 0 at the Equator, to 1 at the Poles. In between the two, the values are distributed according to the function of SIN.

```
W_LATIT_SIN = Sin (D:\earth10k\latitudes / deg)
```

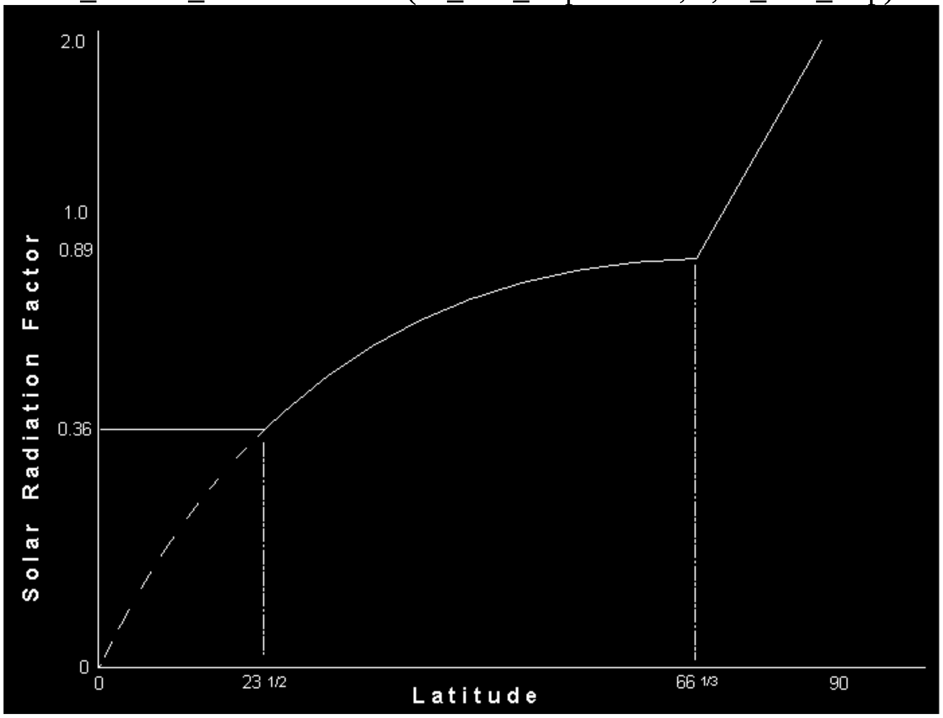
2. Generally, climate is a function of latitude, which strongly correlates with the intensity and duration of sun radiation. The latitudes between the tropic of Cancer and the tropic of Capricorn experience sun radiating from the zenith twice a year. Because of this, it is legitimate to assume that annually these areas receive about the same amount of solar radiation—and for simplification, cloud cover is ignored—and, consequently, create the same impedance while crossing mountain ranges of equal elevation. To reflect this, for all the latitudes between the tropics, the values on the output grid of the SIN function were equalized to the value, which represents the intensity of the solar radiation at the tropics (0.36) on w_latit_sin grid.

```
W_LATIT_TROP = con (w_latit_sin < 0.36, 0.36, w_latit_sin)
```

3. The areas within the polar circles represent harsh living conditions for any humans, forming basically uninhabitable zones referred to as nonecumene. At those latitudes, the sun does not rise at all at least one day per year. To reflect that on the DEM, the values of the elevation grid have been drastically increased to reflect the increasingly deteriorating living conditions toward the poles.

Figure 2-1

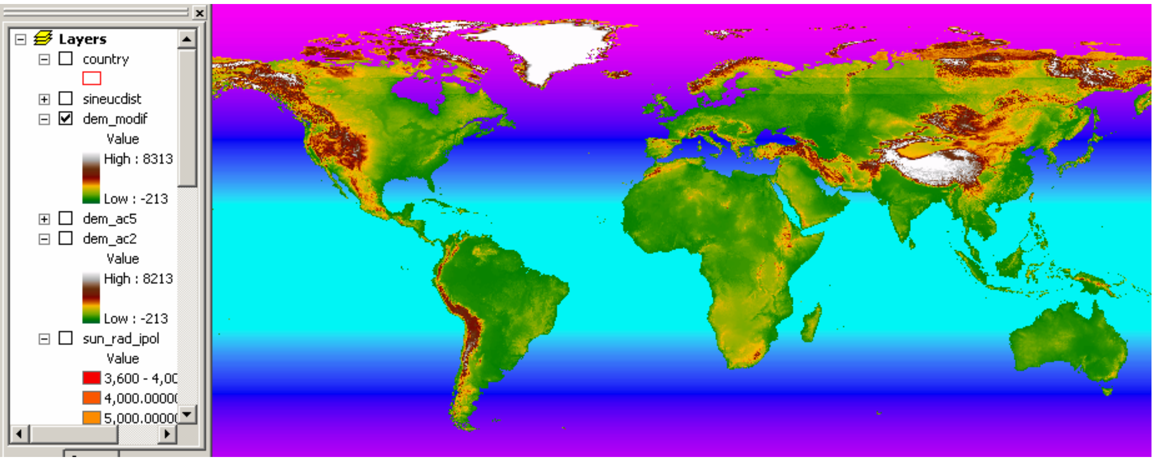
W_LATIT_POLAR = con (w_latit_trop > 0.89, 2, w_latit_trop)



The solid line on the diagram represents the value of the DEM modification factor relative to the latitude.

4. The output of the above CON function was multiplied by the standard DEM. The result of that operation is a modified DEM, where for the low latitudes the elevations are reduced to compensate for warmer climate and more livable conditions. High altitudes at the equatorial zone are more livable and easily penetrable by people compared to the same altitudes located at higher latitudes. The difficulty in inhabiting and migrating through high elevations is almost unbearable beyond the polar circles and increases toward the poles.
```
W_DEM = w_latit_polar * world_dem0
```

Figure 2-2



The modified DEM for the world. The adjusted elevations approximate the relative weight of mountainous barriers. The cyan color on the ocean indicates the tropical zone, while the polar zones are represented by the magenta color.

5. To emphasize the high cost of surviving within or nearby the polar regions and in the highest mountains, which together form nonecumene zones (cells with values greater than 3,000 on the modified DEM layer), the following step was taken:
```
elev3000 = con ( w_dem > 3000, 3000, w_dem)
```

6. To categorize the levels of difficulty of living or migrating through terrain, the raster data created above was reclassified into 10 classes according to the following table:

Class #	"elevation"	weight
1	0–300	0
2	301–600	1
3	601–900	2
4	901–1,200	3
5	1,201–1,500	4
6	1,501–1,800	6
7	1,801–2,100	8
8	2,101–2,400	10
9	2,401–2,700	15
10	2,701–3,000	20

In the resulting data, class number 10 represents nonecumene zone. Classes 9 and 8 approximate territories where living conditions are extremely harsh (called subecumene) for most of a year. Only classes 1 to 7 are considered livable (ecumene). The values of the given weights are proportional to the approximated difficulty of living, with the weights assigned to subecumene and nonecumene zones having prohibitively large values.

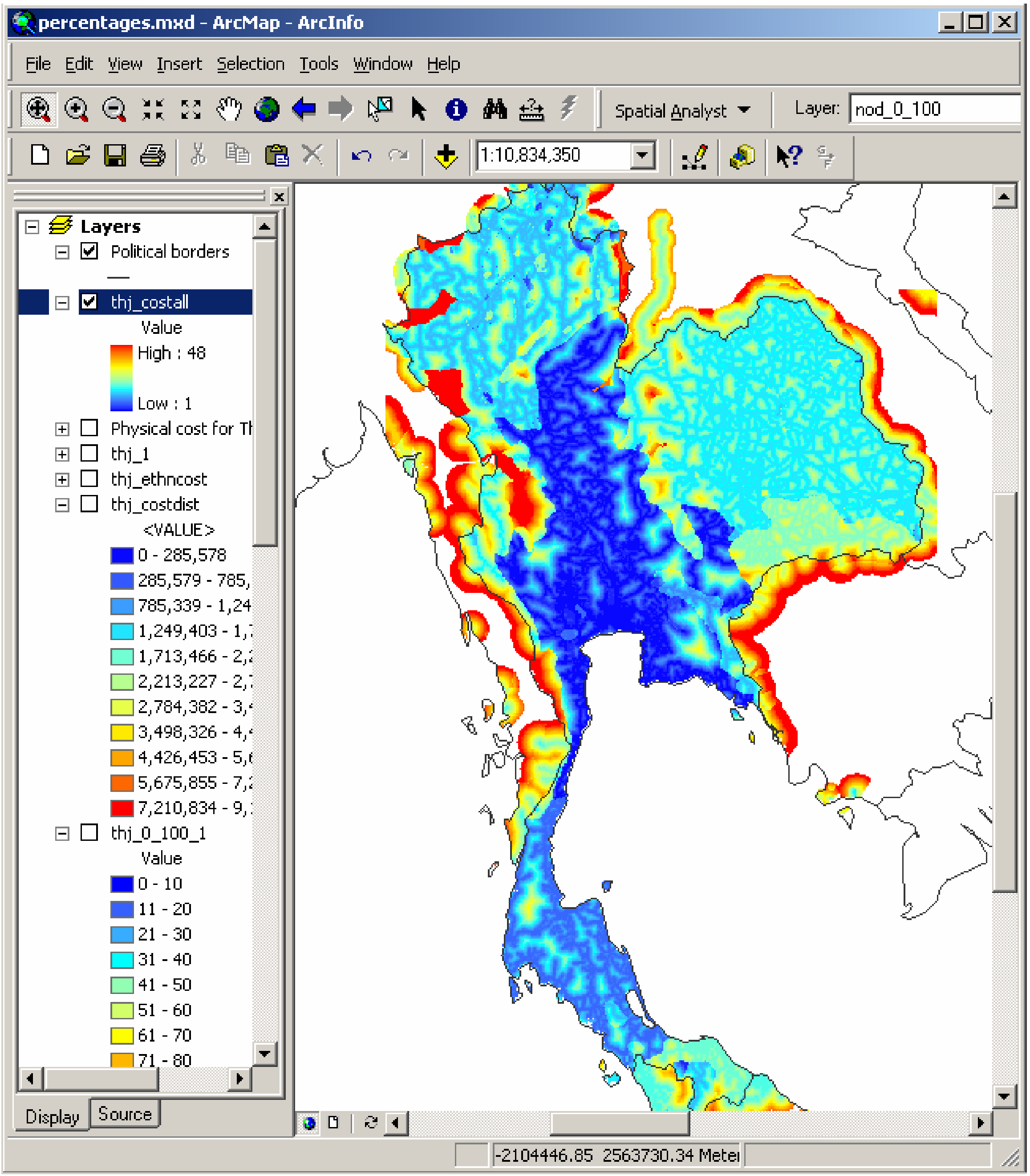
Generally, the farther from the tropics the area is located, the heavier the weight of the elevation component. This way, there is now a normalized elevation in each country applied to the global DEM, in order to compare results for different countries "as apples to apples".

H. The individual components of the physical cost grids D,E,F, and G, were then combined. For grids D,E, and F for every cell (location) the lowest value of the four sources was selected. When these four are merged (using the CON statement to preserve the lowest values present on any of the components), the SLICE operation was performed to produce an output with 20 zones.

I. Now, the G and H grids were then added together into the final PHYSICAL cost surface. The two were combined using the ArcGIS Spatial Analyst grid algebra functionality. The range of the output values of the PHYSICAL grid will typically range from 0 to about 30.

J. To generate the final language diffusion cost grid, the final PHYSICAL cost grid I and the ETHNIC cost grid C were then merged. The merging was performed by adding the two components together. The values of the output integer grid varied from 0 (at the original language territory) to about 50. This grid represents the aggregated cost derivative from all input criteria.

Figure 2-4



Layer: nod_0_100

1:10,834,350

-2104446.85 2563730.34 Meter

The map showing the final cost surface for the Thai Central language in Thailand.

3.0 CostDistance Grid

To generate the CostDistance surface, all three of the following components were generated with this project:

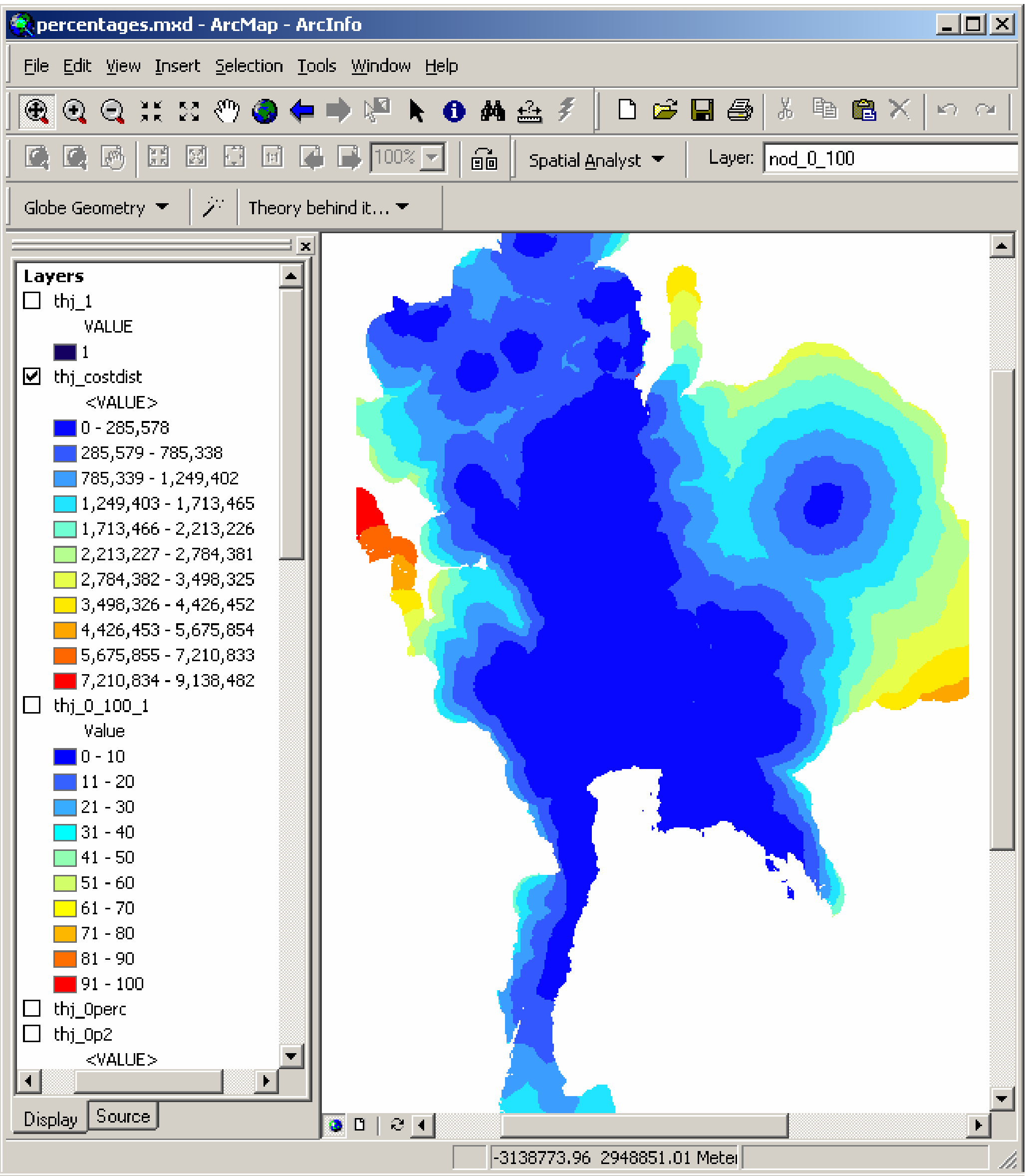
> The final cumulative ethnic and physical cost surface for the given language—(L)
>
> The source grid cells depicting the area where the language is spoken was assigned a numeric value of 1. The remaining areas were all assigned NoData cells <source>
>
> The grid representing the number of people speaking each language (in thousands)—lang_popul grid

K. The CostDistance grid represents the cumulative cost of moving outside the source cells ("b", the area of a studied language) on top of the final cost surface (J). In a way, the COSTDISTANCE function is a variation of the EUCDISTANCE run on the cost surface. Close to the <source>, the values of the CostDistance grid will be low, but they will increase outward according to the final cost surface and the distance from the <source>. On the cost surface input grid (L), the area of the <source> language will have a value of 0.

```
thj_costdist = costdistance (thj_1, thj_costall)
```

For most languages, such as those of Thailand, the maximum values of the output of running the CostDistance function on the final cost surface should fall between about 5 and 10 million.

Figure 3-1


percentages.mxd - ArcMap - ArcInfo
File Edit View Insert Selection Tools Window Help
100%
Spatial Analyst
Layer: nod_0_100
Globe Geometry
Theory behind it...
Display
Source
-3138773.96 2948851.01 Meter

Layers
□ thj_1
VALUE
1
☑ thj_costdist
<VALUE>
0 - 285,578
285,579 - 785,338
785,339 - 1,249,402
1,249,403 - 1,713,465
1,713,466 - 2,213,226
2,213,227 - 2,784,381
2,784,382 - 3,498,325
3,498,326 - 4,426,452
4,426,453 - 5,675,854
5,675,855 - 7,210,833
7,210,834 - 9,138,482
□ thj_0_100_1
Value
0 - 10
11 - 20
21 - 30
31 - 40
41 - 50
51 - 60
61 - 70
71 - 80
81 - 90
91 - 100
□ thj_0perc
□ thj_0p2
<VALUE>

The final CostDistance surface for the Thai Central language in Thailand.

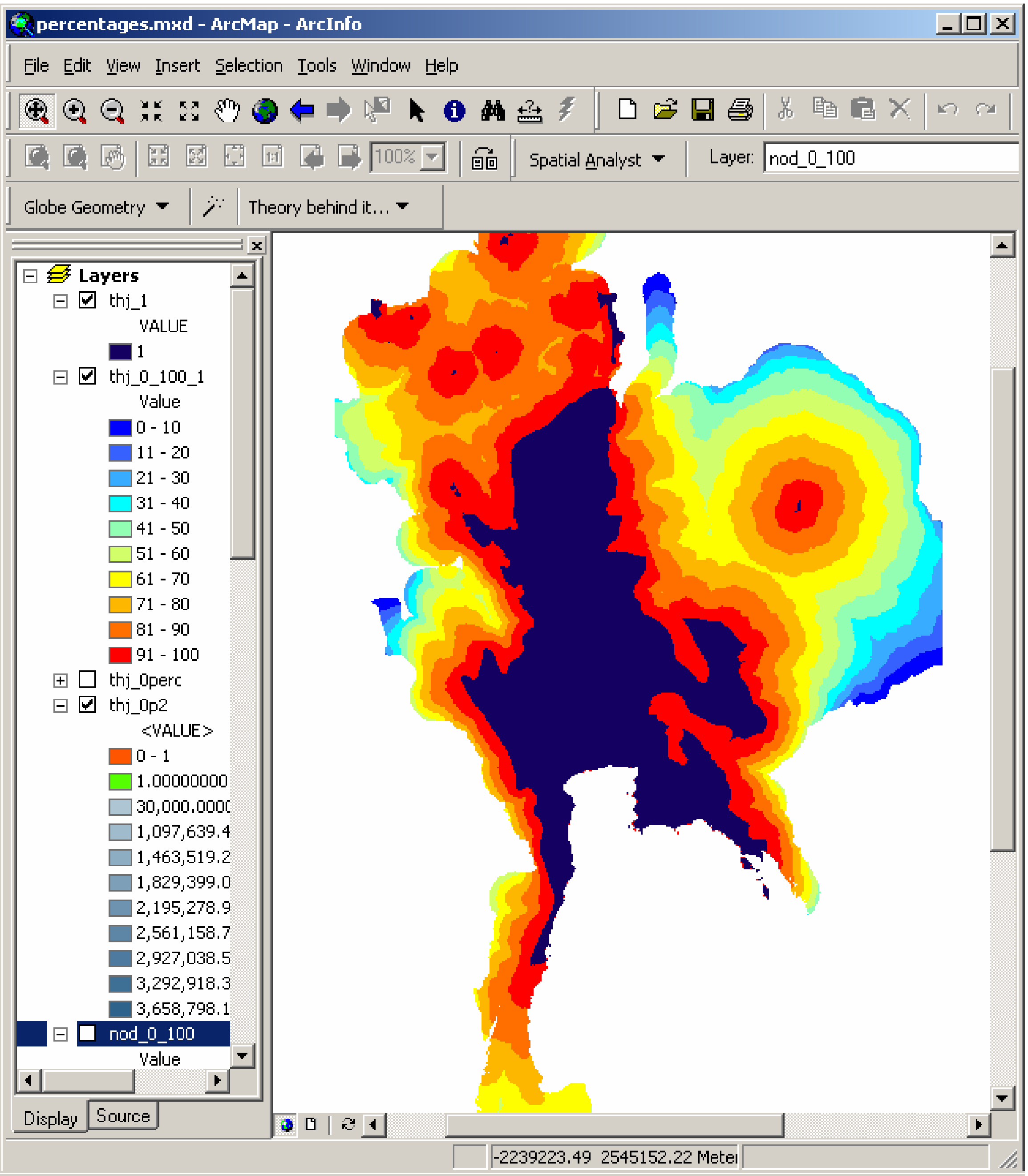
L. The CostDistance grid was expressed in percents of an approximate likelihood of meeting someone speaking the analyzed language. Naturally, the entire area, which represents the language of interest, has a value of 100 percent of the likelihood of finding someone being able to communicate in that language. In order to transform the CostDistance grid (K) into a grid representing a language spatial diffusion expressed in percentages, it is critically important to decide how to determine the 0 percent likelihood. In general, the language's chances to expand are proportional to the number of its current speakers and their position in the country relative to the other ethnic groups (financial and social status of the particular ethnic group, the fertility rate, mortality, literacy rate, languages of TV and radio broadcasting, and the languages in which the elementary education is provided). The heavier the "weight" of the language, the higher the language's chances to expand (based on the above-mentioned factors), and the more widespread it's potential spatial zone of influence.

At this stage, out of the above-mentioned influencing factors, we only had data representing the number of speakers of every language. A global source for demographic factors by ethno-linguistic group is currently unavailable. Thus, the Landscan population density grid was utilized to determine the population of a given language area. This data served as the only determining factor of the language propagation analysis. The number of speakers for a given language is proportional to the size of that language's zone of influence. The approach that was used for this project was based on clipping the output of the CostDistance function at the value being the total number of speakers of that language divided by 10. The probability of 0 percent was assigned to the newly generated "edge" or the outline of the "language territory of influence". In other words, the final CostDistance grid (M) might be seen as a 3D surface (a bowl with irregular slopes). The flat (all 0 cells) "bottom" of the CostDistance surface was defined by the shape of the currently processed language polygon. The steepness and all the irregularities of the slopes going almost indefinitely upward determine the CostDistance surface. The weight of the language (now represented only by the number of the language's speakers) might be envisioned as an amount of a liquid stored in a cylinder. When the liquid is spilled over the 3D CostDistance surface, it will fill it proportionally to the amount of the liquid and according to the slopes of the surface. The contour on the surface reached by the liquid determines the 0 percent probability of the language popularity. The area between the current borders of the language of interest and the outline of that language zone of linguistic influence defines an area representing a language's zone of influence. To generate a grid representing such a zone, the following functions were executed:

```
thj_clip0 = int (con(thj_costdist < (number_of_speakers / 10),
thj_costdist)
thj_0_100 = ((thj_clip0 * -100) / (number_of_speakers / 10) + 100
```

The resulting thj_0_100 grid now had values ranging from 0 to 100. These can be interpreted as the approximated likelihood of a language being spoken in percentages.

Figure 3-2



Thai Central language's "zones of influence" expressed as percentages.

M. In order to view the resulting areas where there is a 70 percent or higher likelihood of the language being spoken the Clip function is used. Clip the language spatial diffusion grid expressed in percents (N) at the 70 percent level.
Thj_70 = con (thj_0_100 > 70, thj_0_100)

4.0 Masking the Uninhabited Territories

The zones of linguistic influences of the studied languages can be considered only for the inhabited areas. At this juncture, the uninhabited areas could be eliminated from further consideration and processing. With the reliable population density data[2] for the entire world of 1 km resolution, it is feasible to come up with a mask covering uninhabited areas. A decision needed to be made at what level of population density the mask needs to be set. In this preliminary study, a value of 3 people/km^2 was used as the default value with an option for the user to change it within the range of 1 to 10, depending on the region of the world.

```
Populated = con (pop_density > 3, 1, 0)
```

As expected, the binary grid resulting from the above operation was fragmented and cartographically unappealing. Raster cleanup is necessary to achieve a cartographically pleasing appearance. The fill/erase functions and methods of the ArcScan extension may be applied here to eliminate clusters of cells smaller than a certain value (e.g., 10). Similarly, the little holes (smaller than five cells) within the large inhabited zones should be filled as well. We have decided not to implement this for the moment.

5.0 Basic Mathematement Notations of the Algorithm

1. *x* and *y*—Geographic coordinates
2. *fxy*—The physical cost of a surface for a cell having coordinates xy, assuming $f_{[xy]}$ is an integer between 0 and 30
3. *exy*—The ethnic cost of a surface for a cell having coordinates xy, assuming $e_{[xy]}$ is an integer between 0 and 30
4. *txy*—The total aggregated value of cost of a surface at a cell having coordinates xy
5. *bxy*—Cumulative cost distance value of a cell having coordinates xy
6. *A*—The number of people speaking the language L
7. *pxy*—Probability of meeting a person speaking the language L in a cell having coordinates xy

To compute the cost distance for a particular cell xy, we have to know the total aggregated cost of the final cost surface for each cell in the territory of influence for a language L. Use the following formula:

$$b_{xy} = \min\{\sum_{i=1}^{n} a_i \ : a_1 = t_{k_0,l_0} = 0,\ a_n = t_{xy},$$
$$\text{if } a_i = t_{k_i,l_i} \text{ then } a_{i+1} = t_{k_{i+1},l_i} \text{ or } a_{i+1} = t_{k_i,l_{i+1}}\}$$

In other words, we connect cell xy with the border of the language L, then we sum the total cost of surface of the cells in this connection. The *bxy* is just the minimum of values of all these connections. The total cost (*txy*) of surface for cell xy is the sum of ethnic and physical costs for cell xy as follows:

$t_{xy} = e_{xy} + f_{xy}$

The formula for *pxy* (depending on *bxy* and *A* is as follows):

$$p_{xy} = \left(\frac{A/10 - b_{xy}}{A/10}\right) * 100\%$$

To avoid negative values we can write *pxy* as follows (then for cells outside the territory of influence, the probability will be zero):

$$p_{xy} = \left(\frac{A/10 - b_{xy} + |A/10 - b_{xy}|}{2A/10}\right) * 100\%$$

Here |*A*| means the absolute value of *A* (e.g., | 5 | = 5 and |- 5| = 5)

6.0 General Recommendations

Data

The DCW roads is not complete for the entire world and the road categories seem to be rather arbitrary. Complete data (all roads, rivers, etc.) should be used in the model in order to generate accurate results.

Updated data (e.g., the most recent and accurate population density or language data) is essential.

More collateral data would enhance the quality and the reliability of the model (e.g., fertility rates and literacy rates for the ethnic groups, languages that the children are taught in local elementary schools, languages in which local TV and radio stations broadcast their programs). Unfortunately the demographic data by ethno-linguistic group is not available at this time. As the model and resulting languages are disseminated we hope that feedback will include this information.

Information on the level of difficulty to cross the international borders was not available and would be arbitrary at best. This would need to be coded or automated in order to be added to the model.

Linguistic proximity data should be developed and made available in a digital format. A diagram showing the distances and levels of linguistic proximity between ethnic groups would make it easier to generate this dataset.

Unanswered Questions (Known Problems)

How do we approach the major languages of the world, such as English, Chinese, Spanish, Arabic, French, and Russian, which are spoken in many regions? How do we show the number of speakers of, for example, English in Nigeria and India, Chinese in Malaysia and the Philippines, or Arabic in Indonesia and Afghanistan, which are countries where these languages are not native, but are commonly spoken?

Cartographic Projection

The problem of choosing the right projection for the project is essential. On one hand, we are aware that the presented model only estimates the spatial diffusion of languages. With this, the problem of any projection's inherent spatial discrepancies is theoretically covered. On the other hand, we want to reduce the potential systematic error related to cartographic projection as much as possible. Consequently, the projection issue indeed becomes critical.

Theoretically speaking, to reduce the inherent error of projection, each continent (or region) should have its own projection minimizing the distance distortion. One of the conic projections (e.g., Lambert or equidistance conic) with the appropriate parameters for every continent should be applied. The problem is that the conic projections work well only at continental scale. These projections are proper for individual countries as big as the USA, but *cannot* be reasonably used for the entire world. Analyzing individual countries based on a world projection is a big compromise (on the accuracy of the frequently used distance parameters).

It would be best to choose from the following:

> Preserving relatively good accuracy of language probability estimation when running the model on the individually projected data for six continents.
>
> Compromising the spatial accuracy of the probability estimate for ease of use and running the model on one of the world projections (e.g., Mollweide or equidistance cylindrical).

When it comes to the world projections, none of them is truly suitable for this study. Still, the Robinson would most likely work best. Because of some technical reasons, raster data projected into Robinson and Mollweide did not appear correct using ArcGIS 8.3 software. Thus, we compromised and settled on equidistant cylindrical projection. The equidistant cylindrical projection look similar to the geographic layout except that the units are meters instead of decimal degrees. It is great at the wide equatorial zone, it is fine for middle latitudes, but the distance error propagates toward the poles.

Final Remarks

We anticipate future ground-truthing to calibrate the design of the model.

The model should use different parameters depending on the geography of the study area to appropriately represent the numerous contributing factors and reasonably share the relative weights among them.

We are fully aware that some major information layers, which play a significant role in linguistic propagation study, were not accessible, while the accuracy and/or completeness of some other layers was questionable.

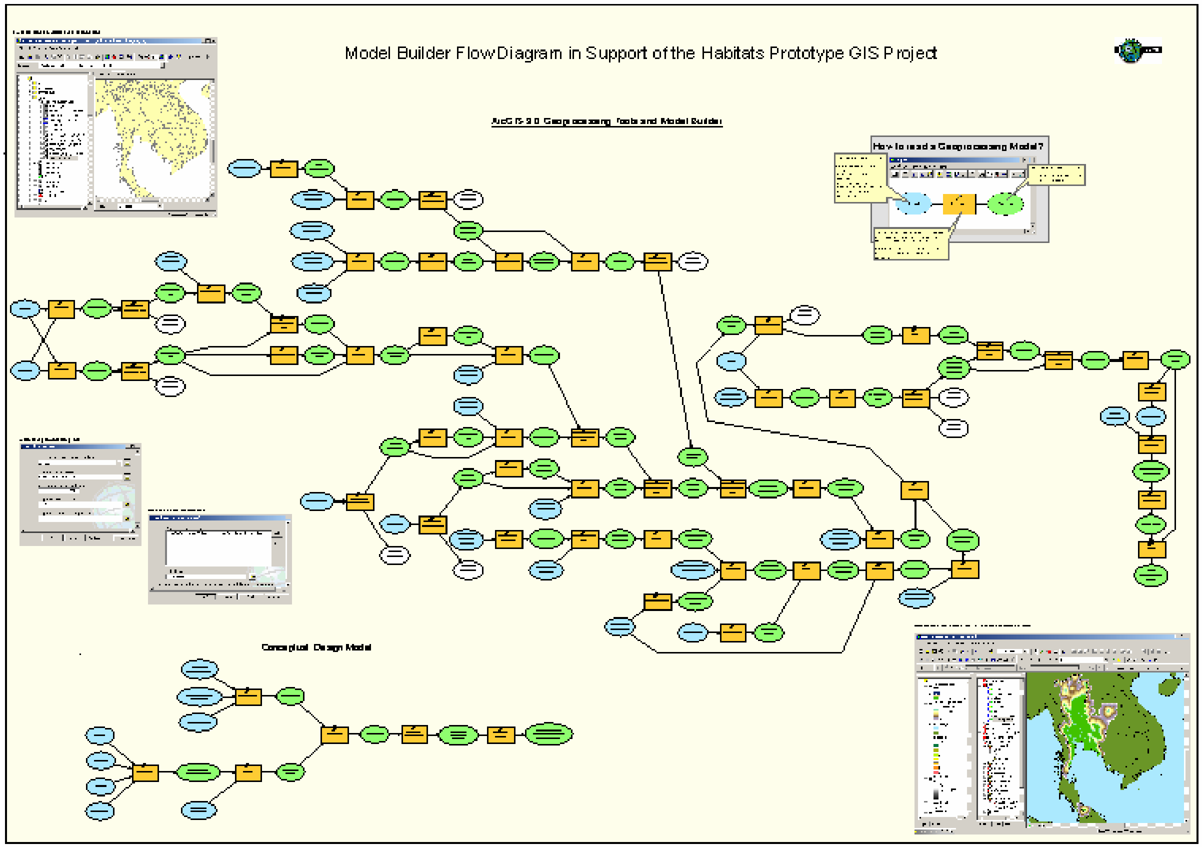
An expert knowledge of the study area will be crucial in achieving reasonable results from running the model and in providing further information necessary to the refining of the model.

GIS seems to be an excellent tool to generate the desired information on linguistic propagation.

Current Development

Currently we are transferring the model to ArcGIS 9 ModelBuilder. Automation of each process is greatly enhanced as well as the ability to assess fine tune each function included in the model.

Figure 6-1


Model Builder FlowDiagram in Support of the Habitats Prototype GIS Project
ArcGIS 9.0 Geoprocessing Tools and Model Builder
How to read a Geoprocessing Model?
Conceptual Design Model