For office use only
T1 _____
T2 _____
T3 _____
T4 _____

Team Control Number
IMMC23287123

Problem Chosen
A

For office use only
F1 _____
F2 _____
F3 _____
F4 _____

2023
(IMMC)
Summary Sheet

Using Land - A Valuable Resource

Summary

The recent push for rapid industrial and social activities have fuelled demands for the optimization of land in rural areas. In general terms, land optimization rests upon two central premises: the geographical features of the land and the differing characteristics of the installed facilities. In order to take effective measures to optimize land use, the government agency of New York State has contacted our team to devise a best metric that outlines the best use of land, determine more than two options for the best metric, evaluate the impact for the installment of a semi-conductor fabrication facility, and apply the model for use in another geographical region other than Syracuse.

Inspired by the **Evaluation Model**, we used the evaluation system that consists 7 parameters for determining the ”best” metric that would evaluate both long-term and short-term benefits. These parameters represent the Economic Profit, Short-term Paid Back, Long-term Sustainability, Construction Cost, Yearly Impact, Environmental Impact, and Community Need. We used the **Analytic Hierarchy Process** with surveys and manual assessments and then used Dimensionless Normalization to process values onto a scale between 0 and 1.

Additionally, we utilize **K-Means Clustering** to find the best 5 infrastructures out of an initial pool of 16 infrastructures. The selected 5 infrastructures are then inputted into the **Genetic Algorithm** which further optimizes the model. Further, each infrastructure choice is compared with an empty option where no infrastructure is installed. This ultimately enabled an optimized city plan, made up of 2 final infrastructures: pig ranch and crop crops. Our model is superior due to its unique feature of elimination from the initial 16 infrastructures: this ensures that the final selected options are objective and to the best interests of the people and the government.

Furthermore, we further tested the robustness of the model by introducing a new semiconductor fabrication facility to our land. We concluded that this lead to an increasing in residential areas and a decrease in ranches to accommodate for the population influx.

Finally, we assessed the robustness of our model by implementing it in the rural areas of **Greenbow, Alabama**, which results in the optimization planning of: pig ranch and residential areas. This result highly conforms to the highly rural landscape of Greenbow, Alabama, which yields predominantly accurate results.

We also discussed the strengths and weaknesses of our model and reached the conclusion that our model is able to consistently produce accurate results, but may be limited in application due to its underlying assumptions.

Keywords: Analytic Hierarchy Process, Genetic Algorithm, K-Means Clustering, Land Optimization, Economic Model

GAO

LETTER TO DECISION MAKERS

Mr. Chairman and Members of the Committee:

At your request, the Office of Special Investigations has gathered information on the optimized land usage of Syracuse city and on the future course of those efforts.

We have determined a quantitative decision metric, crafted upon the pillars of economic development, social conditions, and environmental considerations, which we incorporated through AHP. With the aid of K-Means Clustering and Genetic Algorithms, we are able to devise a city plan by gridding different geographical regions, and ultimately propose a land plan that largely consists of pig ranches and residential areas for an optimized result.

At the same time, law enforcement officials have informed us the construction of a semiconductor fabrication facility north of Syracuse. We have come to the consensus that its construction would increase residential areas while decreasing the areas of ranches.

Finally, the application of the Syracuse model is extended to Greenbow, Alabama, where the we saw a significant increase in pig ranches and crop farms. This highly conforms to the actual situations in Greenbow.

Team IMMC 23287123

041853

1 Introduction

The classical economic theory suggests that the effective use of land is the foundation and cornerstone for a successful and flourishing economy. Taking this theory into account, we aim to develop a mathematical model that optimizes the use of land to benefit sustainable business ventures through a systematic assessment of its environmental impacts, geographical conditions, community needs, and cultural traditions. It has been evidenced that successful entrepreneurial activities can invigorate industrial developments in all sectors, driving economic spurts in surrounding regions. Through continuous investments and ploughed-back profits, citizens and local governments could enjoy the fruits from improved life qualities and increased financial security. To reinforce this aim, we propose the following objectives for our mathematical model.

This paper details an in-depth inspection regarding the city of Syracuse located in New York State, USA as the first testing ground for our model. Geographically, Syracuse is a loosely populated region located at a latitude of 43 /degrees N in the northeastern part of the United States, with ample annual precipitation and adequate water and power supplies. This provides an optimal business environment for establishing a wealth of agricultural, energy, and entertainment facilities. As advisors for Syracuse city, we have considered the following options: an outdoor sports complex, cross-country skiing facility, a crop farm, a grazing farm, a regenerative farm, a solar array, an agrivoltaic farm, and an agritourist center.

Though each option could potentially fuel a range of benefits, we aim to develop a model that dictates an optimized use of land to achieve the maximum economic effect. In doing so, we need to account for a number of consumer factors, including tourism, farming, environmental factors, education, and healthcare. Maximizing the utility of the land, our model also aims to consider the geographical characteristics such as elevation, slope, aspect, tree cover that are specific to the city of Syracuse. This is to build a model that is widely applicable to the real world scenarios.

2 Problem Restatement and Overview of Our Work

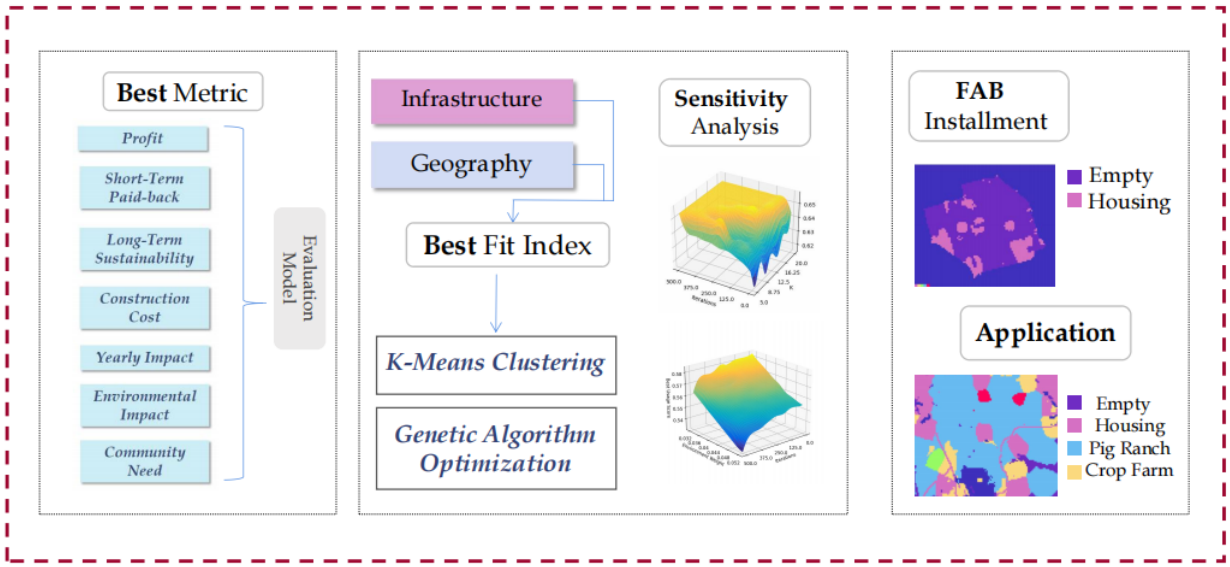
Our overarching goal is to determine an optimizes distribution of land that entails a number of entertainment, agricultural, energy, and tourist centers. To specify, the required tasks and our implementations are as follows:

2.1 Problem Restatement

Task 1 instructs us to dictate a quantitative decision metric that defines a best option for the use of land. Task 2 stipulates the application of at least two of the land use options as mentioned in the introduction and to draw a quantitative analysis on the adaptability of the option. This involves assigning values to each region on the map into grids. In Task 3, we were instructed to determine the impact of the induction of a new factory into our land in Syracuse. We were given the conditions that 9,000 people will be employed who would make an average of 100,000 dollars per year. In Task 4, we would like to extend the application of our model for use in a new environment, which we determined to be Greenbow, Alabama. This would involve changes in certain parameters of our model to suit the new environment.

2.2 Flow Chart of Our Work

The flow chart represents modeling processes to accomplish Task 1, Task 2, Task 3, and Task 4.



BSFA
Framework

3 Assumption and Justification

3.1 General Assumptions

To simply the model and to ensure its adaptability to a wide range of scenarios, we make the following basic assumptions, each of which is properly justified.

Assumption 1: Abnormal weather patterns including droughts, abnormal precipitation, and snow storms can be neglected.

Justification: We assume that the climate data given by the United States Department of Agriculture are reliable and can be applicable to our model for the next ten years. We assume that such changes would not impact the following parameters in our model: tourist traffic, community needs, economic returns, etc.

Assumption 2: The negative environmental damages derived from the construction of facilities in Syracuse can be neglected and would not factor into the environmental damages given in Task 1.

Justification: Though it is evidenced that the construction of public facilities could lead to a certain amount of environmental damage, which may lead to slight changes in patterns such as precipitation and average temperature. However, we render these influences as insignificant. We further contend that these changes associated with construction would not increase nor decrease tourist traffic and community or cultural needs.

Assumption 3: We assume that the residential region in Task 2 and Task 3 is the property of the government and is for rent only.

Justification: Our project is conducted at the request of government agencies and, as a vastly undeveloped landscape, we assume that the residential houses belong to the property of the government. To accommodate a largely itinerant working class, we assume that all houses in the residential areas can be for rent only. This ensures that profits from renting could be generated on a yearly basis, thus ensuring economic sustainability.

3.2 Notations

The notations used in this paper and their relevant descriptions are shown in Table 1.

Table 1 Notations and Descriptions

Notations	Descriptions
P	Annual Net Profit
C	Consumer Population
P_i	Population of the i th city
D_i	Distance from the i th city to the given Syracuse land
W_i	Weight of the i th city
T_P	Payback Period
S	Short Term Benefit Index
T_C	Construction Length
C_I	Construction Length Impact Index
I	Long Term Benefit Index
N_i	Number of Points that Belong to the Cluster of Centroid i
X_i	x Coordinate of Centroid i
Y_i	y Coordinate of Centroid i
$X_{(i,j)}$	x coordinate of the jth grid for centroid i
$Y_{(i,j)}$	y coordinate of the jth grid for centroid i
D_f	Percentage of Forest Destructed
D_w	Percentage of Wetland Destructed
D_c	Percentage of Crop Destructed
E	Environmental Destruction Index
C_n	Community Need Index
B	Best Use of Land Index
K	Number of Clusters in K-means
N_I	Number of Iterations
L_f	Land Fitting Index
d	Distance
x	Power for Distance
P_o	Original Population
P_n	Net Population
C_o	Original Consumer Population
C_n	New Consumer Population

4 Task 1: Best Metric

4.1 Overview

Task 1 requires us to propose a comprehensive quantitative value that is indicative of a "best option" as required by the optimization problem. First, we need to quantify the construction of the infrastructure, establish an evaluation system for the construction, and characterize the impact scores of different parameters. Next, based on the data of different regions, different infrastructural options are applied in the model, and the degree of influence of the construction in different options is obtained.

Thus, we decided to employ an evaluation system based its own nature from the seven indicators: economic profit, Short-term Paid Back, Long-Term Sustainability, construction costs, yearly impact upon construction, environmental destruction, and community needs. These factors not only determine the economic benefits given by the specific infrastructure, but they also take into account a range of social factors that are specific to Syracuse city, which ensures a widely-applicable model.

Referring to the dimension planning evaluation system of the impact of the construction on the city, we here divide the evaluation principle into:

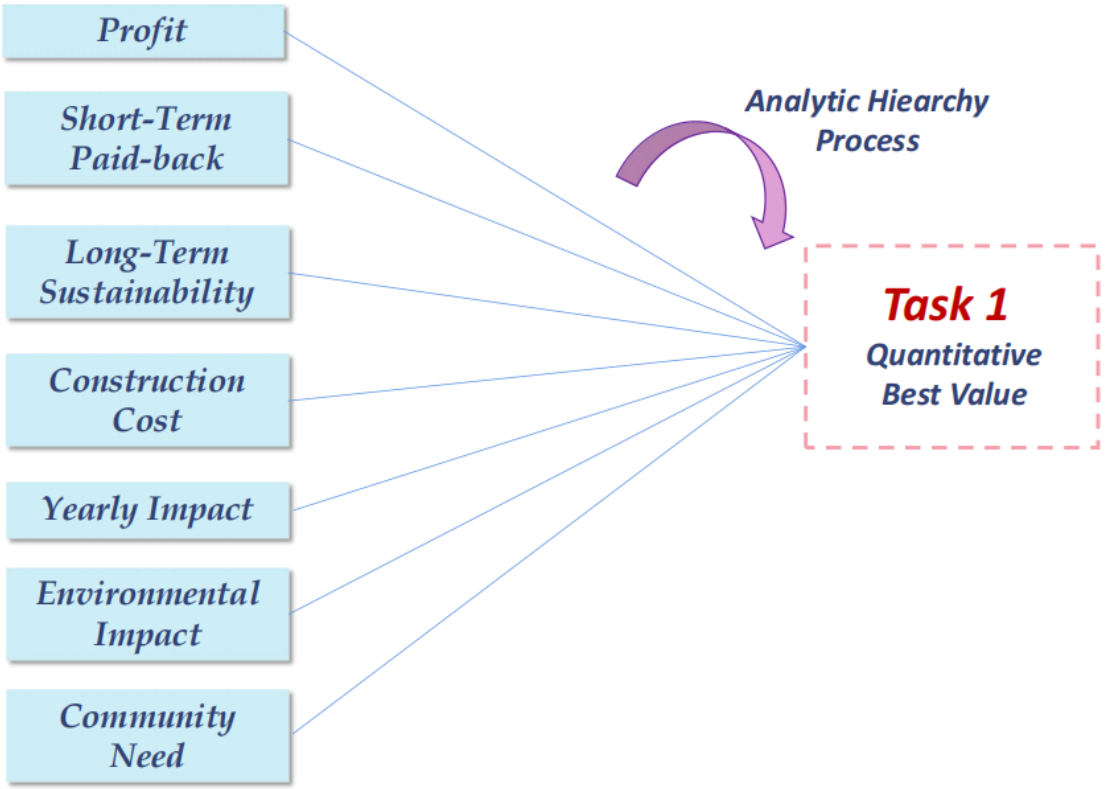


Figure 1 Overview of Task 1

4.2 Factors Affecting Quantitative Best Value

- **Economic Profit**

Refers to the average economic profit gained per year after the installation process is complete over a course of ten years.

To calculate: Consumer spending constitutes the bulk of the annual profit generated by the infrastructure. We factor it in through the equation:

- **Short-Term Paid-Back Time**

Refers to the time in years where the initial construction costs could be paid back.

To calculate: The sigmoid function denotes the short term benefits over installation years. We determined that the sigmoid function generally fits the cost-benefit trend, because, at the beginning of installation, the speed at which construction costs are paid back is relatively small. This would gradually increase significantly following the sigmoid trend, and would then decrease after a certain period of time.

- **Construction Cost**

Refers to the construction costs of all parts of the infrastructure, including personnel, material, and land costs, measured in USD.

To calculate: This data is not calculated but obtained from the reports released by infrastructural companies and within government agencies.

- **Long-Term Sustainability**

Refers to the long-term sustainability regarding the development of the infrastructure. We take this into account because different infrastructures may face different trends in development in the long-term period: while crop farms exhibit long-term sustainability, agritourist clubs may only be demanding for first-time tourists. Over time, their attraction would diminish.

To Calculate: This is determined by an evaluation function which concerns three indicators: the number of tourists in a population as a percentage, the rates at which the public interest has diminished, and the operational cost of the infrastructure, which we assigned the values 0.2, 0.5, and 0.3, respectively. This is due to the fact that public interests regarding the infrastructure play the most pivotal role in determining its sustainability, since consumers are the largest contributors to the infrastructure's profits.

- **Yearly Impact**

Refers to the impact of time (in years) on the "goodness" of the infrastructure, factoring into the degradation of the infrastructural facilities.

To calculate: Previous research have determined of a negative exponent relationship between the yearly construction time and the "goodness" of the infrastructure. We take this result a step further and dictate the following equation founded upon the natural logarithm.

We employ this equation because this equation best fits a multiple array of real life scenarios. Specifically, the "half-life" of a certain infrastructure is around a 2-3 year time span. This equation decreases until, at year 10, the "goodness" of the infrastructure has diminished to a very low level, which is indicative that this infrastructure is out-of-date.

- **Environmental Impacts**

Refers to the environmental impacts caused by the construction of the infrastructure, factoring in the hazardous effects of water, air pollution and deforestation.

To determine the environmental damages caused by the construction of the infrastructure, we propose the following set of criteria.

- **Community Needs**

Refers to the demand from the community-dwellers regarding the specific use of the infrastructure, and how well its suits their daily schedules.

To Calculate: The community need index was defined by three sets of indicators: local culture, education, and diversity. We used tables to denote the impact of these three parameters on the community need index. Local culture, education, and diversity were assigned the values of 0.3, 0.4, and 0.3 respectively. We use a three-point scale due to the fact that users and interviewers respond to three-point scales more effectively than conventional ten-point scales. The community need index is the added sum of all three parameters.

4.3 Using Analytic Hierarchy Process to Weigh the Factors

We considered the followings factors that we want to optimize when we are evaluating whether a land is in its best use: Annual Profit, Payback Index, Construction Cost, Long Term Profit, Development Duration, Environmental Harm Index and Community Fit Index. In order to consider them comprehensively, we need to weigh each factor.

4.3.1 Process of AHP

To determine weights in a math quantity, AHP is often adopted. The main process AHP is explained below.

Construct the Judging Matrix

We use the pairwise comparison and one-nine Saaty method to construct the judging matrix $A = (a_{ji})_{7\times 7}$ which satisfies $a_{ij} \times a_{ji} = 1$ following the equation(1):

$$\begin{array}{ccccccc} 1 & 3 & 4 & 3 & 5 & 7 & 6 \\ \frac{1}{3} & 1 & 2 & 1 & 3 & 5 & 5 \\ \frac{1}{4} & \frac{1}{2} & 1 & \frac{1}{2} & 3 & 4 & 4 \\ \frac{1}{3} & 1 & 2 & 1 & 3 & 5 & 5 \\ \frac{1}{5} & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 1 & 3 & 2 \\ \frac{1}{7} & \frac{1}{5} & \frac{1}{4} & \frac{1}{5} & \frac{1}{3} & 1 & \frac{1}{2} \\ \frac{1}{6} & \frac{1}{5} & \frac{1}{4} & \frac{1}{5} & \frac{1}{2} & 2 & 1 \end{array} \tag{1}$$

Calculate the Eigenvalues and Eigenvectors

The greatest eigenvalue of matrix A is 7.257, and the corresponding eigenvector is $u = (u_1, \ldots, u_n)^T$. Then we normalize u by equation (2):

$$\omega_i = \frac{u_i}{\sum_{j=1}^{n} u_j} \tag{2}$$

Conduct the Consistency Check

The indicator of the consistency check formula is

$$CI = \frac{\lambda_{\max} - n}{n-1}, CR = \frac{CI}{RI} \tag{3}$$

where n denotes the dimension of the matrix. CR is the expression of the consistency ratio. It is used in robust analysis, when the value of $CR = 0.030$ less than 0.1, we conclude that the model is dependable.

4.3.2 Final Solution

As some of the factors are negative factors that we would like to minimize, they would be given a negative sign in calculation. The result of the AHP model, which would be the final equation of calculating the Best Use of Land index, is shown in equation (4):

$$\begin{aligned} index =& 0.379 * AnnualProfit + 0.179 * PaybackIndex - \\ & 0.122 * ConstructionCost + 0.179 * LongTermProfitIndex + \\ & 0.066 * DevelopmentDurationIndex - 0.032 * EnviromentalHarmIndex + \\ & 0.042 * CommunityFitIndex \end{aligned} \tag{4}$$

4.4 Dimensionless Normalization

Since the data from each factor has different quantities and units, it is impossible to analyze them together as a composite index before processing them onto the same scale, in this case, 0 to 1.

Among the seven factors we have considered, the Payback Index and the Construction Cost Index are already within the range of 0 to 1. The Long Term Profit Index, the Environmental Index, and the Community Fit Index all have a fixed scoring range, and thus we could directly scale them equally so that their upper limit would become 1. Last but not least, the profit and the cost of the building would have a much more irregular distribution, and thus we could not directly normalize them. In this case, we ran the program and counted the frequencies of profit and cost at different levels, and the distributions are as presented in the figures.

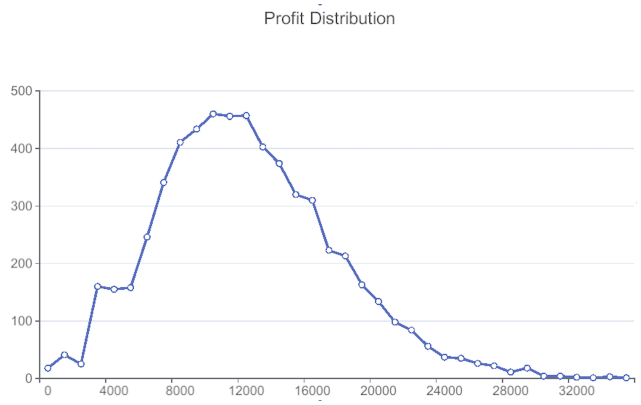


Figure 2 Profit Distribution

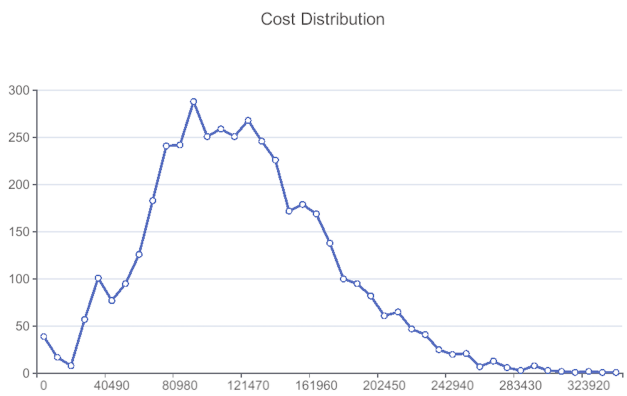


Figure 3 Cost Distribution

We counted the minimum value, lower quartile, median, upper quartile, and maximum value of the results and iterated the program until these values had converged, in this case, it means that they remained the same in further iterations when considering three significant figures. We then collect these data for calculation. Finally, the numbers from the minimum value up to the lower quartile are scaled proportionally to align with the interval from 0 to 0.25. The numbers from the lower quartile up to the upper quartile are scaled proportionally to align with the interval from 0.25 to 0.75. The numbers from the upper quartile up to the maximum value are scaled proportionally to align with the interval from 0.75 to 1. This completes the dimensionless normalization of cost and profit.

5 Task 2: Best Option with Data-Driven Metrics

5.1 Overview

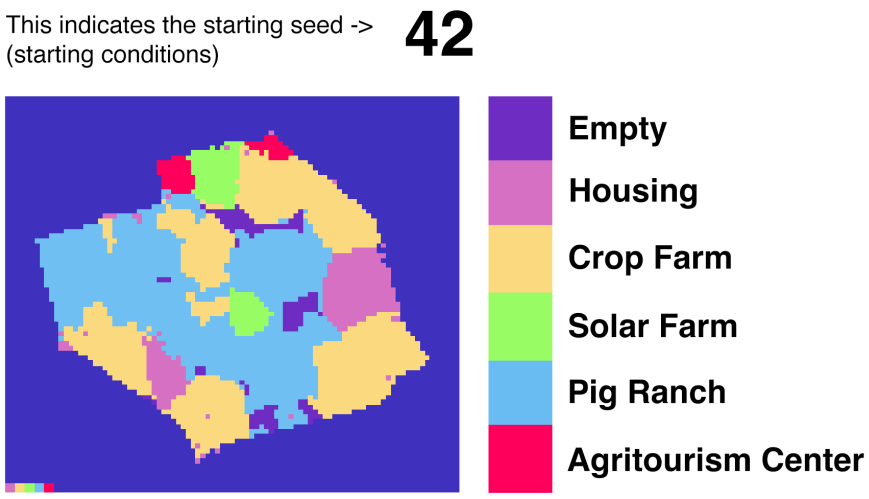


Figure 4 An example of a converging K-means cluster with color keys.

In Task 2, we used a K-means clustering model to find the best option for allocating land in Syracuse. To cluster different regions of the land, we divided the property into smaller grids of land

based on the highest resolution mappings of land cover data. Each pixel corresponds to one unit of $\frac{\text{land}}{\text{grid}}$, roughly equating to 909 m^2.

This model would partition each grid of land to K clusters, resulting in K buildings since each centroid can be considered as an individual building with position and building type. We approach K-means by clustering each grid to the centroid (building) with the highest score, which evaluates Distance and the Fitness Index. The Fitness Index evaluates how "fitting" a specific building would be on a specific terrain type.

The K-means model will show preference to certain building types, reflected by the area of land clustered to these buildings. These buildings have larger areas due to them having overall a higher score, which means they are more suitable for this property. We utilize this relationship to first eliminate certain buildings by removing the building types with the least average clustered areas. This estimation using K-means significantly reduces additional calculations for these buildings for final evaluations, and narrows down the building choices.

The Best Use of Land index, as formulated in Task 1, will be calculated for the top 5 buildings, selected based on highest area clustered on average out of 16 candidates. A Best Use of Land index for any layout of the 5 buildings can then be calculated.

Thus, to find the Best Option, we simply find the layout with the highest Best Use of Land index. The K-means model is shown to converge given a set of initial conditions (initial centroids). We can hence use a Genetic Algorithm that finds the best set of initial conditions resulting in the highest Best Use of Land index to find the best overall option. The genetic algorithm is also shown to converge, thus, we have derived a method that would find the "global best option".

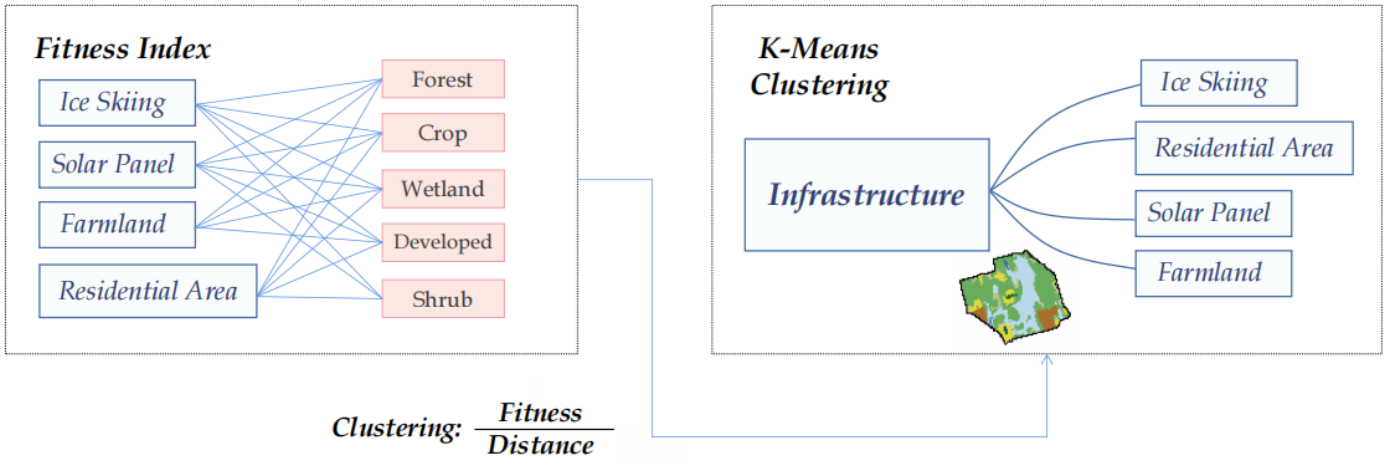


Figure 5 Overview of Task 2

5.2 Land-Fitting Level

Land Fitting Index denotes the level of fitness between a geographical landscape and a specific infrastructure that we wish to install.

To optimize land use, it is essential to evaluate the level of fitness between a certain landscape and the infrastructure which resides on the land. For instance, a wet and humid region would not be the best suit for a solar array but would benefit the installation of a crop farm or an agritourist center. With this in mind, we determined several indicators that would evaluate the "fitness" between

the geographical conditions of the land and the infrastructural facilities for all possible installment combinations.

5.2.1 Land Type

According to the United States Geological Survey, the major types of land in northeastern United States could be defined as the following: forest, wetland, shrub, and plateau. Taking into account of the actual conditions in Syracuse, we decided to modify the framework by adding two additional layers to the existing categories, defined by: **forest, cropland, wetland, and developed.** Shrubs are less common in Syracuse, so we neglect its impact on the final model. Specifically, forest and wetland indicates the natural state of the land where it has not yet been developed by industrialization. Additionally, the labels "croplands" and "developed" signify that the land had been largely developed and exploited for agricultural, industrial, or living purposes.

This is important for the fitness evaluation between the land and the infrastructure, because underdeveloped land require higher transformation expenses, covering expenses such as logging, installing electricity, and transporting life necessities for human use, whereas already developed land do not require such expenses. The physical condition of the land is complementary to the different requirements as demanded by different facilities. Farmlands, on one hand, do not require intensive industrialization to be implemented into use, whereas skiing centers require a high degree of industrialization to account for emergency medications, external communication, and food import.

From the United States Geological Survey, we obtained the map of the small piece of land, and we generalized the given map into the following four geographical regions.

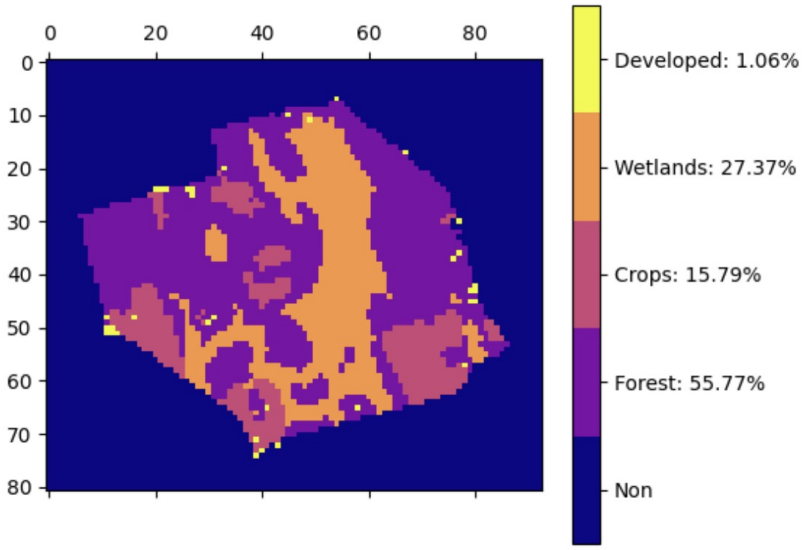


Figure 6 Land Type

5.2.2 Scoring Criteria

To quantify the criteria, we generate several expressions to display them numerically.

- **Criterion 1: Transformation Capacity**
  Transformation capacity refers to the amount of change needed to build a certain type of build-

ing on one of the specified land types. We take the reciprocal of the the percentage of transformation and scale it down by a factor of 2, so that the final transformation capacity would be a value on the 0-5 scale. Data for the percentage transformation of land is obtained via the Multi-Resolution Land Characteristics Consortium from the United States Geology Survey.

$$\text{Transformation Capacity } = \frac{1}{\text{Percentage of Transformation} \times 2} \tag{5}$$

$$0 \leq \text{Transformation Capacity} \leq 5$$

- **Criterion 2: Transformation Difficulty**
  Transformation Difficulty, on the other hand, is a collected variable that draws reference to the amount of equipment, manual labor, and time it takes for a certain geographical land to transform to the specified infrastructure. The transformation difficulty score directly correlates to the easiness of transformation.

  We use a 0-2 scale to denote the transformation difficulty. This is because a 0-2 scale embodies less weight compared to the transformation capacity scale in Criteria 1, which has a scale of 0-5.

Transformation Difficulty		
Relatively Easy	Difficult	Very Difficult
2	1	0

Table 2 Transformation Difficulty Index

- **Criterion 3: Soil Fitting Index**
  The soil fitting level refers to the ”fitness” between the infrastructure and humidity of the soil. This provides a differentiating contrast between wetland and mountain ranges, and such differences should be taken into consideration when evaluating the ultimate land fitness level.

  The soil fitting index is directly correlated to the fitness between the soil and the infrastructure. A score of 2 represents the highest score while a score of 0 indicates the lowest score. This is conducted through manual assessment that evaluates the humidity of the soil based on geographical characteristics obtained from the Multi-Resolution Land Characteristics Consortium from the United States Geology Survey.

Soil Fitting Index		
High Fitting	Basic Fitting	Low Fitting
2	1	0

Table 3 Soil Fitting Index

- **Criterion 4: Climate Fitting Index**
  Climate fitting level refers to the level of fitness of the natural environment to the infrastructure. This is majorly determined by the percentage of woods that make up each portion of the land. Satellite images and numerical data regarding the distribution of woods can be obtained through Google Map and the government website from the United States Geological Survey. Climate Fitting Level is directly correlated to the amount of fitness, categorized by a 0-2 scale.

Climate Fitting Index		
High Fitting	Basic Fitting	Low Fitting
2	1	0

Table 4 Climate Fitting Index

5.2.3 Quantifying Total Score

We quantify the above indicators by assigning objective values to each land condition and the type of infrastructural facility, as expressed in the tables above. In the tables below, the vertical columns of the graph denotes the potential infrastructure installments; the horizontal rows present the five types of landscapes: forest, cropland, wetland, developed, and shrub. For the first criterion, we use the 0-to-5 scale to define the fitness level and for criteria 2 to 4, we use the 0-to-2 scale, as the following tables suggests.

We ultimately obtain the following tables, each grid containing the fitness value of each landscape to its corresponding infrastructure.

	Forest	Crop	Wetland	Developed
Residential Area	4.714	5.833	5.296	7.667
Farmland	4.625	11.000	6.214	2.625
Skiing Center	3.556	3.000	4.000	1.714
Solar Array	4.500	4.000	5.657	0.500
Outdoor Sports	4.214	5.833	3.269	7.667
Ranch	5.125	6.500	6.500	2.125
Agritourist Center	4.125	7.667	3.500	4.000
Shopping Mall	3.556	3.125	4.250	11.000
Gymnasium	3.214	4.125	3.000	8.833
Restaurant Plaza	3.625	4.500	2.714	7.500
Hotel	3.556	3.526	3.000	7.500
K-12 School	2.500	3.625	2.667	10.5
Karoke Disco	4.214	2.714	1.769	9.333
Coffee Library	4.125	3.000	1.714	8.500
Park	4.929	7.250	3.500	3.714
Amusement Park	3.510	3.556	1.714	4.556

Table 5 Land Fitting Index for 16 Infrastructures

5.2.4 Best Value

To quantify the Best Value as obtained in Task 1, we collected data from a collation of government websites. The data we use mainly includes the official data of various U.S. Government departments, many of which can be directly reported through industry reports, such as the cost of various types of residential areas, the profit gained from ranches and farmlands, and yearly construction expenses in American cities and towns.

Database Names	Database Websites	Data Type
USDA	https://www.ers.usda.gov/	Economic Research
EnergySage	https://news.energysage.com/	Energy Data
Home Advisor	https://www.homeadvisor.com/	Local Home
Trust for Public Land	https://www.tpl.org/	Land Economy
University of California Cooperative Extension	https://ucanr.edu/sites/ucanr/County_Offices/	Agriculture and Natural Resources
Google Scholar	https://scholar.google.com/	Academic Paper

Figure 7 Data Source Collation

5.3 Short-listing

5.3.1 Data Processing

We firstly acquired the official surface cover map produced by the U.S. government online. We then classified the surface cover map into several major categories that were pertinent to the problem statement, including forest, wetland, developed, and crop. However, as discovered in observation, the current proportions of land differed from those given in the problem statement, with forest covering 56%, crop covering 16%, wetland covering 27%, and developed covering 1%.

5.3.2 Using K-Means

To partition the land, we grid the map to turn the problem from a continuous nature to a discrete nature, such that each pixel represented one unit, resulting in a total of 3299 units. We employed a K-means clustering algorithm to classify the land into distinct groups based on shared characteristics. This is achieved through the following steps.

Step 1: Initialization of centroids

We randomly defined K points and assigned each of them a building type, ensuring that the coordinates of the points did not overlap.

Step 2: Reclassification of points

We calculated the score using the formulae $\frac{\textbf{Land Fitting Index}}{\textbf{distance}^K}$ and assigned the highest score to the corresponding cluster center. Land Fitting Index refers to the level of fitness between the land and the specific infrastructure to be installed. The data for the Land Fitting Index is displayed in Table 5, and is obtained through manual evaluation.

Step 3: Update of centroids

We then recalculated the position of the cluster center by taking the average of the x and y coordinates of all points within the cluster.(average coordinates, need mathematical expression)

Step 4: Iterations until convergence

We repeated steps 1-3 until the cluster center positions have converged.

5.3.3 Choosing the Five Most-favored Infrastructures

Finally, we evaluated the 16 different building types by calculating the size of the area for which each building type was assigned by using K-means Clustering. Our aim is to find the best 5 infrastructures which would indicate the best suited construction options, selected from an initial pool of 16 initial infrastructures. This is to eliminate the unsuited building options which do not fit the Syracuse landscape and to ensure an ultimately optimized solution that takes into account of all factors.

The 16 infrastructures are: outdoor sports facility, ranch, agritourism center, shopping mall, residential area, farmland, ice skiing center, solar array, gymnasium, restaurant plaza, hotel, K-12 schools, Karaoke disco hall, library, park, and amusement park.

Our criteria is as follows: the building with the largest area size was deemed the most suitable one for construction in the area. We adjusted the parameter K and ran about 100 iterations for each value of K to obtain the rankings of the whole area. Interestingly, we found that the rankings converged and were consistent for each K.

Results: Our statistical analysis revealed that the 5 buildings (among the 16 pre-selected buildings) that were most favored for construction on the land were: **farm, pasture, housing, solar array, and agritourism center**.

Figure 8 Five Most-favored Infrastructures

5.4 Using Genetic Algorithm to optimize the solution

A given set of initial conditions, or the position and building type of K clusters, will converge to a best layout as seen in Fig. 10. This shows that the K-means clustering model will find a best layout, given a set of buildings. This can be seen as the local optimum to our problem. To find the potential global optimum, we will utilize a Genetic Algorithm that tries to find the best set of initial

conditions (best combination and position of buildings) that yields the highest Best Use of Land index as formulated in Task 1.

5.4.1 Flowchart and Pseudocode of Genetic Algorithm

Algorithm 1: Genetic Algorithm

Input: Instance Ω, size α of population, rate β of elitism, rate γ of mutation, number δ of iterations, the total population P

Output: Solution X

```
// Initialization
Generate α feasible solutions randomly, each with 12-parameter long genes;
Save them in the original P ;
for i ← 1 to δ do
    // Elitism based selection
    n_e = α * β;
    Select the best n_e solutions in P as P_1;
    // Crossover
    n_c = (α − n_e)/2 ;
    for j ← 1 to n_c do
        Select X_A and X_B from P;
        Generate X_C and X_D from X_A, X_B;
        P_2 ← {P_2} ∪ {X_C, X_D};
    // Mutation
    for j ← 1 to n_c do
        Select X_j from P_2;
        Mutate each bit of X_j by rate γ and get X'_j;
        Check(X'_j);
        Update X_j with X'_j in P_2;
    // Updating
    P ← P_1 + P_2;
return the best solution X in P
```

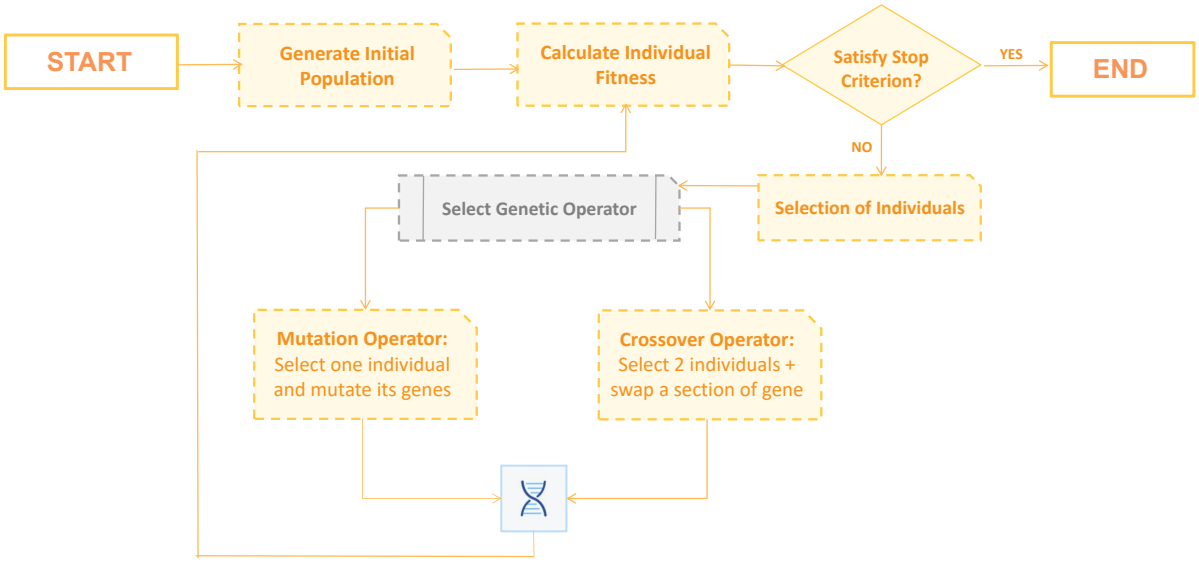


Figure 9 Flowchart of Genetic Algorithm

5.4.2 Process of Genetic Algorithm

Genetic Algorithm is divided in to several phases to obtain the final result. The specific steps are listed as follows.

Step 1: Initialization of Population (Coding)

In this genetic algorithm, a gene is represented by a centroid in the solution. We define a chromosome as a set of centroids. We create 100 chromosomes by randomly selecting K centroids for each chromosome, with the x and y coordinates and the building type all randomly assigned. Hence, the population is a collection of chromosomes.

Step 2: K-Means Processing and Evaluation with Metrics

We run the K-means model formualted earlier in the paper until the cluster has converged, and then we calculate the Best Use of Land index according to the metrics defined in Task 1.

Step 3: Selection

Inspired by the Darwinian principle "Survival of the Fittest", we establish our main gain to locate the region where more optimal solutions can be obtained, given that there should be a balance between exploration and exploitation of the search space. The Fitness proportionate selection is used, also known as roulette wheel selection, as a genetic operator used in Genetic Algorithms to select potentially useful recombination solutions. The selection process is based on the chromosomes's calculated score from its K-means cluster. 50% of the chromosomes, or 50 sets of centroids with the lowest scores will be eliminated. This vacancy will be filled through reproduction of the top 50 chromosomes.

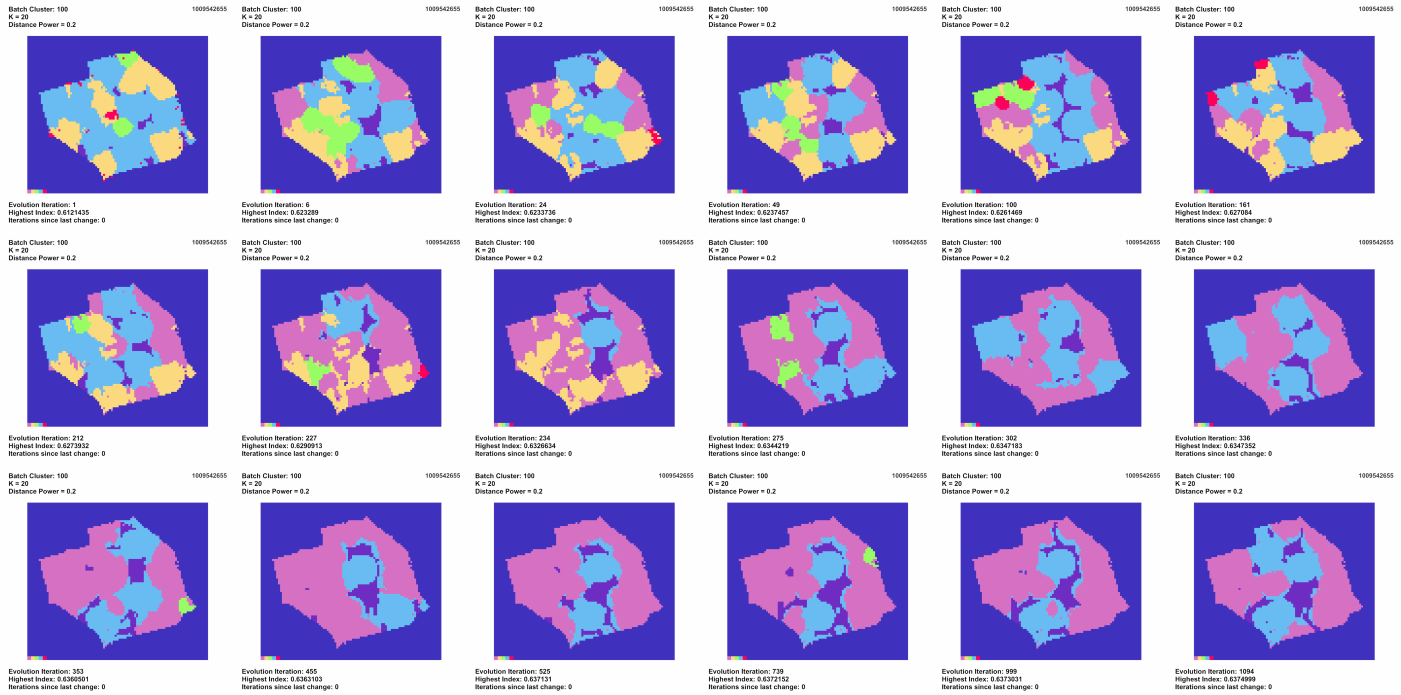


Figure 10 Sequence of images showing the convergence of a global optimum using the Genetic Algorithm.

Step 4: Reproduction

a) Crossover

Crossover is the most vital stage in the Genetic Algorithm, which is acting on behalf of a population group. For each vacant chromosome, we randomly select two distinct chromosomes from the

top 50. We combine certain genes (centroid) of each chromosome to create a new chromosome of same length, which would be the offspring that will fill this vacancy.

b) Mutation

In a few new offspring formed, some of their genes can be subjected to a low random probability mutation. The newly generated chromosomes undergo further changes. Each centroid has a 50% chance of experiencing a displacement in its x and y coordinates as well as a reassignment of the building type. Mutation happens to take care of diversity among the population and stop premature convergence, as well as finding potentially fitter chromosomes.

Step 5: Iterations until Convergence

The process would be iterated until there is no improvement in the maximum score, which the solution is said to converge. This would produce the final globally optimized result that is gained from a combination of K-Means Clustering and Genetic Algorithms.

5.5 Result

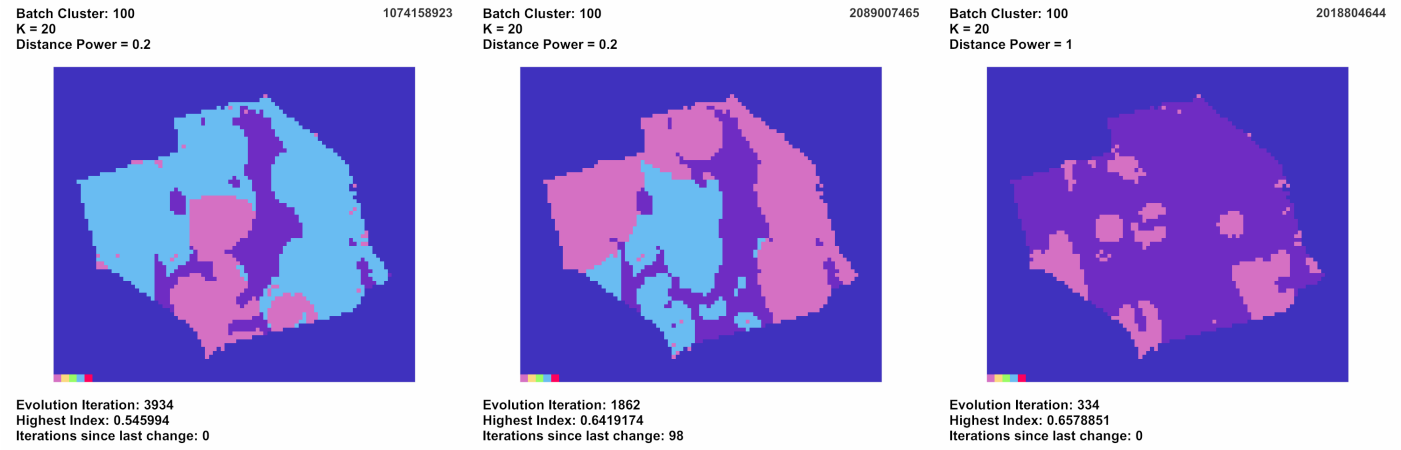


Figure 11 Different Optimization Result

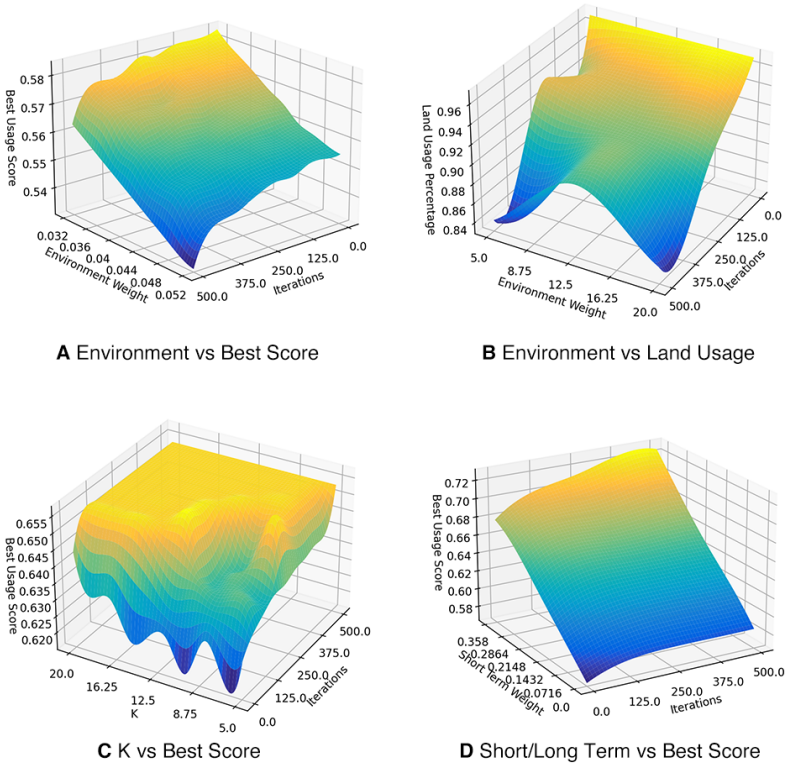
Our results of the genetic algorithms for optimized solutions are presented in the figure, which indicates a general trend of convergence over 3,934 iterations. The pink, dominating regions on the graph indicates the residential area option, a result that is largely logical since housing unites yield a high annual profit. Additionally, the blue regions refer to the pig ranch, and the yellow regions resemble crop farms. Finally, the purple regions are areas which are left empty due to conservational concerns. The result in highly expected given the rural situation in this specified piece of land, as other options, such as amusement parks and shopping malls, would not be profitable.

5.6 Sensitivity Analysis

In sensitivity analysis, we varied the the iteration number, the weight for Environmental Harm Index, the balance between weight of Payback Index and weight of Long Term Profit Index (short-long term relationship), and the K for K-means.

For all diagrams, the Good Use of Land index increases with the iteration number significantly, proving that our genetic algorithm is stably leading the division of land to a better score. Figure C could also show the relationship between K and the scores of the solutions. As K increase, the division of land is more flexible due to more clusters, and thus there would be more potential for the Genetic Algorithm to improve the performance of the land division.

In terms of weight parameters, the Good Use of Land index is more sensitive to the change of short-long term relationship than to the change of weight of Environmental Harm index. On the other hand, figure [refer to the figure about land usage] could also clearly shows that the land usage would decrease as the weight of Environmental Harm index increase, this is due to the fact that the more land it used, the more harm it would cause, and thus the Genetic Algorithm would guide the result to less land usage.



A Environment vs Best Score

B Environment vs Land Usage

C K vs Best Score

D Short/Long Term vs Best Score

Figure 12 Sensitivity Analysis

6 Task 3: Influence of Fab Installation

6.1 Impact Modeling

6.1.1 Population Impact

To model the population impact of the new factory, we assumed that 60% of the new employments in Clay would be immigrants from outside of New York State. We then used a Gaussian Function to model the declining relationship between the distance and the amount of population that would travel to our land. We set the function to give a weight of 1 when the distance from the land is 0 and a weight of 0.05 when the distance is 100000 meters, and the formula used was:

$$W_i = e^{-(\frac{D_i}{5.778\times10^4})^2} \tag{6}$$

6.1.2 Population Increase Calculation:

Using the formula, we calculated the original weighted population, which was the potential consumer population, to be 188108 from the 16 major cities near the land. After the immigration caused by the new factory, the consumer population increased by 13.5%, reaching 213,580.

$$C = \sum P_i * W_i \tag{7}$$

6.1.3 Impact on Projects:

The increase in the consumer population would have a positive impact on some projects. Firstly, the vacancy of the house unit would disappear because the housing units that are originally 90% rent out could be fully filled, leading to an annual profit level of 223.8 dollars per unit square meter. Additionally, the agritourism center would experience an increase in profits by the same percentage increase of the consumer population, which is 13.5%, leading to annual profit levels of 2.8047.

6.2 Result

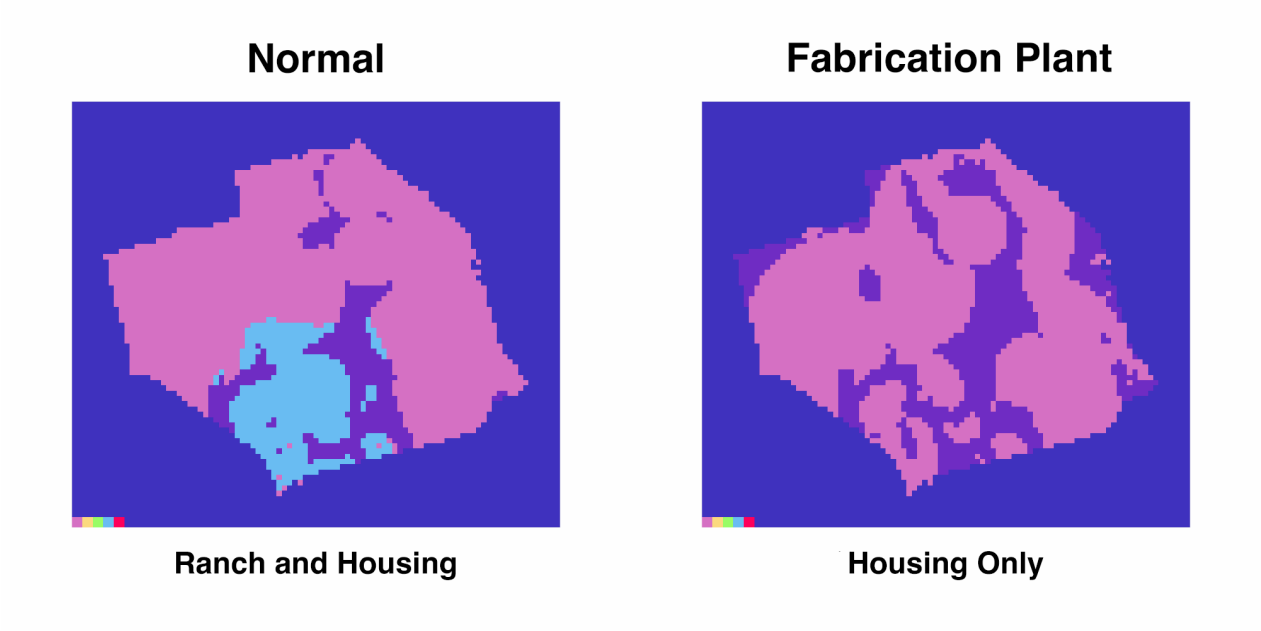


Figure 13 Changes to the Best Use of Land layout after the Factory Installation

We can modify our model from Task 2 by changing the scores of the buildings that are effected by this new factory installation, namely prices of housing and tourism income. We observe the change in our model by giving the same seeds to our random initialization, which results in the same starting conditions. Thus we can observe the effects of new weights in a controlled environment.

The converging results at 500 mutations show a preference to constructing additional housing units instead of ranches. As the increase in housing profits would increase both the short-term gain and the long-term gain, increasing the score of the fabrication plant. During this, we also saw an increase in the frequency of agritourism centers. This however was still not as profitable / best as constructing additional housing units.

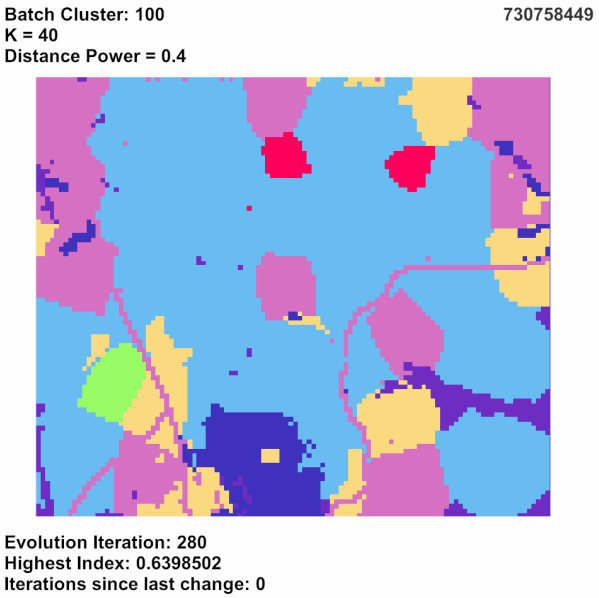
In addition, there is a higher percentage of area left for environmental preservation, mostly around the forest and wetland regions. This shows that the model used some of the additional

profits gained from housing units to compensate for additional conservation of natural habitats. Which is due to a result of our environmental harm index.

7 Task 4: Generalization of Our Model

To test the robustness of our model, we applied our model to a rural region in Greenbow, Alabama. We first obtained the geographical landcape of Greenbow from the USDA website.

Greenbow, Alabama is a rural area primarily dominated by forests and undeveloped regions. It does not have significantly high mountains nor wetlands. In the results that we have generated through K-Means Clustering and Genetic Algorithm in Task 2, we concluded the results in Figure 16: the majority of the land is clustered into a pig ranch (blue), while the pink regions indicate residential areas. This result largely conforms to the actual situations in Greenbow, Alabama, where the majority of its people dwell in farms, which demonstrates the accuracy of our model.



8 Conclusion

To conclude, our model implements the Analytic Hierarchy Process to calculate the ”best” metric that encompasses seven parameters, and then applies AHP to assign parameters. Additionally, we utilize K-Means Clustering, coupled with Genetic Algorithm, to locate the optimization plan for Syracuse. Furthermore, we further tested the robustness of the model by introducing a new semiconductor fabrication facility to our land. Lastly, we also applied the same model to fit Greenbow, Alabama’s landscape, which yields highly accurate results.

Strengths

We devised an AHP Model, K-Means Clustering Model, and Genetic Algorithms to find the optimized solution for land use, which takes into account of most situations comprehensively. Through sensitivity and robustness analysis, we have proved the high stability and strength of our model.

Future Improvements

We have only incorporated four land types: developed, wetland, forest, and croplands. This is not comprehensive of all the existing land types across the United States. In the future, we aim to improve the adaptability of the model by incorporating land types such as deserts and shrubs.

References

[1] "U.S. Organic Production, Markets, Consumers, and Policy, 2000-21." USDA ERS - Home, https://www.ers.usda.gov/.

[2] Fields, Spencer. "EnergySage Blog." EnergySage Blog, 10 Mar. 2023, news.energysage.com.

[3] —. "HomeAdvisor - Find Local Home Repair and Improvement Services." HomeAdvisor, www.homeadvisor.com.

[4] "Trust for Public Land: Connecting Everyone to the Outdoors." Trust for Public Land, 2 Mar. 2023, www.tpl.org.

[5] University of California, Division of Agriculture and Natural Resources. "About UC Cooperative Extension." © 2023 Regents of the University of California, ucanr.edu/sites/ucanr/County Offices.