Policy Extraction for Wildfire Resilience: A PIRS-Based Approach

Bernette Chan
Cal Poly San Luis Obispo
bchan42@calpoly.edu

Jun Noh
Cal Poly San Luis Obispo
jnoh02@calpoly.edu

Nidhi Shinde
Cal Poly San Luis Obispo
nashinde@calpoly.edu

Zarina Yunis
Cal Poly San Luis Obispo
zyunis@calpoly.edu

***Abstract*—Wildfires continue to grow in frequency and intensity across California, requiring legislators to write policies that enhance wildfire resilience. Counties have several plans with conflicting and wordy policies, making it difficult for city planners to evaluate the efficacy of these initiatives. This report explores natural language processing (NLP) techniques used to identify policies related to wildfires. By applying methods like rule-based text extraction and Large Language Model (LLM) prompting, we extract policy statements from four Atascadero County plans. These policies are classified into categories using topic modeling. The ultimate goal of this project is to partially automate a system that scores each policy based on its potential to improve wildfire resilience. Using geographic information systems (GIS), we can map scored policies spatially so that counties can prioritize areas with a high risk of fire-related damage. Moving forward, we aim to extract policies across more counties and refine our categorization strategy to begin the scoring and evaluation process.**

I. Introduction

Wildfires have become an increasingly severe and frequent threat across the United States, especially in the Western states. Over the past 30 years, the number, size, and intensity of wildfires have risen dramatically due to both natural and human-made causes. Climate change has led to hotter temperatures, prolonged droughts, and increased fuel availability, while urban expansion into fire-prone areas and human activities such as power line failures and accidental ignitions have further exacerbated the problem. According to the National Interagency Coordination Center, the number of acres burned annually has steadily increased, causing significant economic, environmental, and human losses.

In response to this growing threat, policymakers and urban planners must integrate wildfire resilience strategies into land-use planning and hazard mitigation policies. However, a major challenge in this process is the inconsistent and fragmented documentation of these policies. Many local governments struggle with poorly formatted planning documents, conflicting hazard mitigation strategies, and a lack of standardized methods for evaluating wildfire policies. This lack of cohesion makes it difficult for city planners to systematically assess and improve wildfire resilience efforts.

To address this issue, we apply data science techniques to assist city planners in identifying fire-related policies within their planning documents. By analyzing policy language and categorizing relevant content, our approach allows planners to efficiently mark policies within the Planning Integration for Resilience Scorecard (PIRS). This tool provides a structured framework for evaluating how well wildfire resilience is integrated into existing plans, ultimately helping cities strengthen their wildfire preparation and response strategies.

II. Dataset Description & Feature Engineering

A. Dataset

The data was provided to us in the form of four large city planning documents:

- **Atascadero General Plan (GP)** (PDF) – A long-term city development plan with policies organized by element chapters.
- **SLO County Multi-Jurisdictional Hazard Mitigation Plan** (PDF) – A document outlining strategies for reducing natural disaster risks, with fire-related policies dispersed throughout sections.
- **Community Wildfire Protection Plan (CWPP)** (PDF) – A shorter document (25 pages) focusing specifically on wildfire policies, some presented in tables.
- **Atascadero Climate Action Plan (CAP)** (Word Document) – A structured climate policy document, often formatted in tables for clarity.

Three of the documents were PDFs and one was a Word document, with each one having hundreds or even thousands of pages. Due to the varying nature of each city planning document, we initially had trouble creating a pipeline to pinpoint and extract policies in a way that was consistent across all four documents. Some of the documents had their policies cleanly listed with a bold heading with the measure number, measure title, and subsequently the measure text. This structure allowed for us to iterate through each page of the document and use regular expressions to directly extract only the relevant policies and then reformat them nicely into a dataframe. While this technique was successful among the more structured documents, we quickly realized that it fell short extracting policies in the documents where policies were not as explicitly defined. This made it difficult to use the same python libraries and direct text extraction techniques to pull each policy. To mitigate this issue, we implemented a couple of different techniques.

B. Our Approach

Two of the methods we used to find policies were table extraction and Large Language Model (LLM) approaches. Upon examining the documents, particularly the City of Atascadero Final Climate Action Plan, we discovered that most of the relevant policies were organized into tables with categories for the measure label, actions, and more. We used this structure to split the relevant text into its respective tables. After coming to this realization, we also removed extra sections from the text that were unrelated to policy, and solely focused on the pages with policy-related text. This made our extraction method much more time-efficient because the text data we were examining was on a much smaller scale. Using the Pandas library in Python, we stored cleaned policies and their labels (number and letter combinations) in a dataframe. After finding relevant policy tables from each page in a document, we wrote the extracted tables to new csv files, each containing a policy number and a policy description. While some of the documents' policies were easy to extract, other documents had policies with varying formats, which made it difficult to implement only one technique to pull policies, goals, and measures from the text. For example, the Wildfire Protection Plan had policies both in tabular format and also integrated into the paragraphs of text.

III. Attempted Policy Extraction Approaches

A. *Word2Vec*

Word2Vec, or word embedding, is a text extraction technique that represents words as numeric vectors to capture their semantic meaning. It is useful for comparisons, with closer vectors representing greater similarity between words. We used Word2Vec as a preliminary approach to categorize policies as fire-related, which was our ultimate goal. We hoped to find meaningful words that were adjacent to the word "fire", which we could then use to search for relevant policies. The two Python libraries we used to accomplish this task were Natural Language Toolkit (NLTK) and Gensim. NLTK contains many text preprocessing libraries for tasks like tokenization, which breaks down text into small fragments that are easier for a machine to understand. There are multiple options for tokenizing words, including by word, character, and subword, from which we chose to tokenize by word. Additionally, we manually labeled sections of the text as wildfire-related or non-wildfire related so we could discern how the Word2Vec model returned words in close proximity to "fire." We hoped that we would see insightful differences between the two types of sections, which would help us identify fire-related policies later on. After processing the text data, we fit our cleaned text to the Word2Vec model in the Gensim library to find the seven most similar words to the word "fire." However, upon running the model, we found that the results were not insightful. The words in our Word2Vec model were fairly unanimous across the fire-related and non-fire related sections, so we shifted gears to other policy-extraction approaches.

B. *Part of Speech Tagging*

Part of Speech (POS) tagging is a less popular, but promising approach to identifying actionable policies. Text analysis libraries like SpaCy have functions that label the grammatical category of each word in a piece of text (noun, verb, adverb, etc). Since most policies listed begin with a capitalized, present-tense verb, this method would ideally capture a lot of implementation actions and general goals. We tested POS tagging on the Atascadero Climate Action Plan (CAP) because the document is relatively short with just 266 pages. SpaCy's tagging function successfully extracted at least one actionable policy per page; however, many common verbs were ignored due to issues with spacing, special characters, and misidentification of certain verbs as other parts of speech. Splitting the document into smaller chunks and using different text-processing libraries did not improve accuracy. If this approach yields better results in the future, we can consider using POS tagging to filter out text that LLMs like Gemini mislabel as policies.

C. *Named Entity Recognition*

Named Entity Recognition (NER) is a natural language processing technique used to identify and classify entities within text into predefined categories such as persons, organizations, locations, and dates. It is commonly used for extracting structured information from unstructured text, which is highly applicable to our dataset of extracted policies from differently structured documents.

We used spaCy's pre-trained NER model to extract entities from our documents, starting with the Atascadero General Plan. This approach included:

1) Processing text data from the document using spaCy's NLP pipeline.
2) Identifying named entities and categorizing them into predefined labels (e.g., `ORGANIZATION`, `DATE`, etc.).
3) Saving extracted entities for analysis to determine if they could help identify any fire-related policies.

However, the NER results showed that while the model successfully recognized general entities such as locations, organizations, and dates, it did not identify specific policy-related terms.

This model could be improved by tuning a custom NER model. Specifically, we could define policy-related entities such as "Housing Policy" or "Fire Mitigation Policy" within the documents. The topics extracted from LDA could assist in refining this model. Additionally, NER can be used to classify policies as fire-related or non-fire-related by identifying key entities within the text.

IV. Current Policy Extraction Techniques

To preprocess the data, we split the documents into structured and unstructured. This meant documents with policies that are easily identified through tables, specific headings and formatting could be classified as structured. The Atascadero General Plan and Final Climate Plan were structured documents since they had a very formatted and organized way to display their policies. Since they could be easily identified, we used a method of **rule-based text extraction** to pull out these policies. However, unstructured documents had policies sparsely distributed among the text. This included the Hazard

Mitigation and CWPP, which had policies in many different formats and in different forms like from tables as well as embedded throughout the text, even to the point where some policies were implicitly stated. For unstructured documents, we utilized a **large language model** to extract the policies for us.

A. Rule-Based Extraction

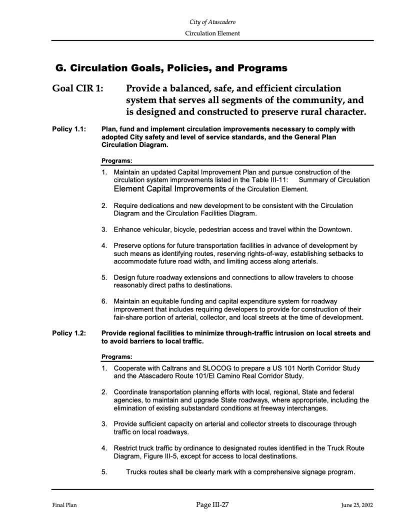
Rule-based text extraction involves chunking each document line by line to identify which text is a policy given regular expressions and other structure-based logic. The two structured documents we were given, the General Plan and Final Plan, had certain emphasized (bolded, italicized, larger font, etc.) words to signify that it was a start of a policy or other action items. Each document had a slightly different format, so it was necessary to alter the code to extract each of the policies pertaining to different documents.

The General Plan is organized by element chapters like Housing Element, Safety Element, Land Use Element, Circulation Element (see Figure 1) etc. Each chapter contains sections with structured goals, policies, and programs.

- **Goals** are always capitalized and included an element abbreviation (e.g., Goal LOC 1 for the Land Use, Open Space, and Conservation Element, and Goal SAF 2 for the Safety and Noise Element).
- **Policies** are listed under each goal (e.g., Policy 1.1, Policy 1.2).
- **Programs** are listed under each policy as a numbered list.

To extract this structured data, we used regular expressions to identify and pull out formatted goal, policy, and program names along with their corresponding text. The challenge was determining where each section ended, which was typically marked by extra spacing.

Each extracted item was stored in a hierarchical dictionary to maintain relationships between goals, policies, and programs. This structure was later converted into a spreadsheet for the client's review (see Figure 2).



City of Atascadero
Circulation Element

G. Circulation Goals, Policies, and Programs

Goal CIR 1: **Provide a balanced, safe, and efficient circulation system that serves all segments of the community, and is designed and constructed to preserve rural character.**

Policy 1.1: **Plan, fund and implement circulation improvements necessary to comply with adopted City safety and level of service standards, and the General Plan Circulation Diagram.**

Programs:

1. Maintain an updated Capital Improvement Plan and pursue construction of the circulation system improvements listed in the Table III-11: Summary of Circulation Element Capital Improvements of the Circulation Element.
2. Require dedications and new development to be consistent with the Circulation Diagram and the Circulation Facilities Diagram.
3. Enhance vehicular, bicycle, pedestrian access and travel within the Downtown.
4. Preserve options for future transportation facilities in advance of development by such means as identifying routes, reserving rights-of-way, establishing setbacks to accommodate future road width, and limiting access along arterials.
5. Design future roadway extensions and connections to allow travelers to choose reasonably direct paths to destinations.
6. Maintain an equitable funding and capital expenditure system for roadway improvement that includes requiring developers to provide for construction of their fair-share portion of arterial, collector, and local streets at the time of development.

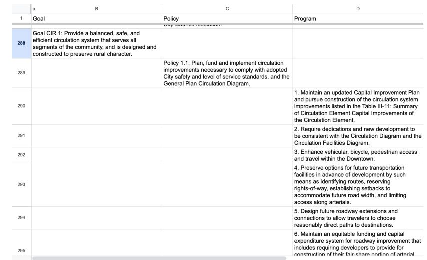
Policy 1.2: **Provide regional facilities to minimize through-traffic intrusion on local streets and to avoid barriers to local traffic.**

Programs:

1. Cooperate with Caltrans and SLOCOG to prepare a US 101 North Corridor Study and the Atascadero Route 101/El Camino Real Corridor Study.
2. Coordinate transportation planning efforts with local, regional, State and federal agencies, to maintain and upgrade State roadways, where appropriate, including the elimination of existing substandard conditions at freeway interchanges.
3. Provide sufficient capacity on arterial and collector streets to discourage through traffic on local roadways.
4. Restrict truck traffic by ordinance to designated routes identified in the Truck Route Diagram, Figure III-5, except for access to local destinations.
5. Trucks routes shall be clearly mark with a comprehensive signage program.

Final Plan Page III-27 June 25, 2002

Fig. 1. Example of a General Plan document structure.



	B	C	D
1	Goal	Policy	Program
288	Goal CIR 1: Provide a balanced, safe, and efficient circulation system that serves all segments of the community, and is designed and constructed to preserve rural character.		
289		Policy 1.1: Plan, fund and implement circulation improvements necessary to comply with adopted City safety and level of service standards, and the General Plan Circulation Diagram.	
290			1. Maintain an updated Capital Improvement Plan and pursue construction of the circulation system improvements listed in the Table III-11: Summary of Circulation Element Capital Improvements of the Circulation Element.
291			2. Require dedications and new development to be consistent with the Circulation Diagram and the Circulation Facilities Diagram.
292			3. Enhance vehicular, bicycle, pedestrian access and travel within the Downtown.
293			4. Preserve options for future transportation facilities in advance of development by such means as identifying routes, reserving rights-of-way, establishing setbacks to accommodate future road width, and limiting access along arterials.
294			5. Design future roadway extensions and connections to allow travelers to choose reasonably direct paths to destinations.
295			6. Maintain an equitable funding and capital expenditure system for roadway improvement that includes requiring developers to provide for

Fig. 2. Extracted General Plan policies in spreadsheet format.

A very similar approach with direct extraction is used for the Final Plan. This document includes measures and actions. These items were distributed throughout the document, but it was also identified throughout a table, which seemed to be the best approach. We parsed through the table as the measure and action items with regular expressions since they were numbered well.

B. Querying LLMs

Unstructured plans posed a unique challenge for extraction because policies were either listed out or embedded within paragraphs. Large language models (LLMs) like Google's Gemini and OpenAI's ChatGPT have the ability to extract specific answers from a piece of text. These models are black-box technologies, so we can't analyze the model architecture or parameters to understand the reasoning behind the mistakes it could make. For instance, we could ask Gemini to extract all actionable policies from a single page; this prompt could work perfectly on one page and still miss certain policies on another page. To maximize accuracy, we wrote queries with the goal of capturing as many potential policies as possible, even if some were irrelevant for our use case.

Our process involves three steps: splitting the document into chunks, querying the Gemini API for each chunk, and saving these responses in a text file.

1) Chunking: We split each plan by pages since the free tier of Gemini's API has 2 restrictions - daily limits on token (character) processing and a limit of 15 queries per minute, or 1,500 queries per day. These limits prevented us from splitting each plan into chunks smaller than a single page if we were handling a large document. For example, the SLO County Hazard Mitigation Plan contains nearly 1,500 pages, so if we were to query more than one document of this length, we would have to wait several days for our results. Splitting by pages was also convenient for error checking since we could go through each page and count the number of policies extracted by the LLM. The exception to this rule is for small documents like Atascadero's Community Wildfire Protection Plan (CWPP), which is 25 pages long.

2) Querying: We then queried Gemini's API with each chunk, asking the model to extract policies and return them in their original format. Our initial queries were more generalizable but failed to deliver consistent results for each chapter of a plan. Since smaller documents (less than 500 pages) are split by paragraph and larger documents are split by page, we developed distinct queries for each document type.

- **Large Documents:**
  1) *Structured documents* (clearly listed goals, policies, and programs)
     `"Goals are typically listed with abbreviations, followed by policies, each of which has associated programs. Extract the goals, policies, and programs from this paragraph. For each policy, include the exact wording. If a policy has associated programs, extract those as well."`
  2) *Unstructured documents*
     `"Extract policies from this piece of text. If you encounter a policy, do not paraphrase it. Keep the exact wording."`
- **Small Documents:**
  `"Extract both explicit and implicit policies from this text. A policy can be a rule, guideline, or a recommended action. Provide the exact wording."`

We tested the query for small documents on the CWPP and the first query for large documents on the Hazard Mitigation Plan. The CWPP has paragraphs with implied policy actions, so this prompt was able to extract all explicit and implicit policies. The Hazard Mitigation Plan was partially structured with goals, policies, and programs, so the first query helped the LLM identify those patterns easily.

3) Saving Responses and Error Checking: Once these policies were extracted for each chunk, they were saved in csv files and exported to spreadsheets. We measured accuracy by counting the number of policies the LLM successfully extracted and dividing that count by the total number of policies for that document. We flagged chunks that did not capture policies and manually examined those paragraphs, which helped us optimize our queries to extract as many policies as possible. As a result, Gemini's API captured all policies in the CWPP. Error checking has not been completed for any other documents.

The extracted policies from both rule-based and LLM-based methods are compiled in a spreadsheet.[1]

V. Policy Categorization: Latent Dirichlet Allocation

A. Motivation

After extracting policies from the planning documents, the next challenge was categorizing them into meaningful topics. Given the large volume of extracted text, manually sorting policies into relevant themes would have been inefficient and subjective. To address this, we implemented Latent Dirichlet Allocation (LDA), a topic modeling technique that helps uncover hidden themes in a collection of documents. LDA assumes that each document consists of a mixture of topics, where each topic is represented by a distribution of words. Once we obtained the distribution of words for each topic, we manually assigned labels to the topics based on their most relevant keywords, ensuring that they aligned with wildfire resilience themes. This additional step helped refine the automated categorization and improve interpretability. By leveraging this probabilistic approach, we enabled a more systematic classification of policies, making it easier for city planners to analyze and interpret wildfire-related policies. To validate our approach, we compared the extracted topic probabilities and assigned topics to the PIRS scorecard, which served as our ground truth since these were the policies our clients had already labeled as relevant.

While our dataset included four documents—the Atascadero General Plan, the SLO County Multi-Jurisdictional Hazard Mitigation Plan, the Community Wildfire Protection Plan (CWPP), and the Atascadero Final Climate Action Plan (CAP)—we primarily focused our LDA analysis on the Atascadero General Plan. This document contained well-structured policies with dense content, making it an ideal starting point for topic modeling before expanding to the other documents.

B. Initial LDA Model and Refinements

Our first LDA model treated each policy as a separate document, allowing us to analyze how individual policies aligned with different topics. We initially set the number of topics to 15, as this provided the best alignment with the PIRS scorecard. This approach offered valuable insights but also revealed certain limitations. Due to the density of the policies, which contained multiple programs and action items, the model sometimes failed to assign a dominant topic to a given policy. Instead, it distributed low probability values across multiple topics, with some policies receiving probabilities as low as 0.067 across all 15 topics, making classification less meaningful. To improve topic assignment, we adopted a more granular approach by treating each program within a policy as a separate document. This method allowed LDA to work with more fine-grained text units, capturing important distinctions between different action items and ensuring a more precise topic classification. Specifically, we segmented policies into individual programs and increased our number of topics to 16 to account for the additional documents created by this segmentation. After refining our approach, the updated LDA model was able to map relevant topics to each policy/program more effectively. Each policy/program had a clear, dominant topic with a high probability.

C. Results & Analysis

Following the implementation of LDA with both 15 and 16 topics, we evaluated the extracted topics and their probability distributions to measure the model's effectiveness. Below is a list of the topics identified for each model, with the most relevant topics to the PIRS scorecard highlighted:

[1]Available here

TABLE I
15-TOPIC MODEL

Topic	Topic Category
1	Community Development & Risk Management
2	Environmental Hazards & Anti-Discrimination Policies
3	Urban & Residential Zoning
4	Downtown Planning & Design Standards
5	Historic Preservation & Zoning
6	Natural Resource Conservation & Emergency Management
7	Housing Development & Smart Growth
8	Parks, Trails, & Public Spaces
9	Tourism & Real Estate Policy
10	Agricultural & Geologic Land Use
11	Park & Water Resource Planning
12	Infrastructure & Public Facilities
13	Transit & Disaster Preparedness
14	Waste Management & Public Services
15	Mixed-Use & Commercial Development

TABLE II
16-TOPIC MODEL

Topic	Topic Category
1	Building Accessibility & Design Regulations
2	Housing Policy & Neighborhood Compatibility
3	Residential & Mixed-Use Development
4	Land Use & Public Facilities Planning
5	Environmental & Archaeological Considerations in Development
6	Transportation & Traffic Management
7	Zoning & Aesthetic Standards
8	Noise Regulations & Land Use
9	Master Planning & Infrastructure Development
10	Economic Development & Housing Goals
11	Emergency Preparedness & Land Use
12	Open Space & Recreation Planning
13	Historic Preservation & Community Identity
14	Education & Sustainable Growth
15	Transportation & Parking Standards
16	Environmental Conservation & Emergency Response

We analyzed the probability values assigned to each policy/program to determine how well the model captured dominant themes. In the 15-topic model, some policies were spread too thin across multiple topics, leading to lower probability values (e.g., 0.067) across all topics, making classification less distinct. However, with the transition to 16 topics, each policy/program was more strongly associated with a single dominant topic, with probability values ranging from 0.4 to 0.8, leading to clearer topic assignments.

To illustrate the impact of our refinements, we include a comparison showing how the LDA model initially struggled to identify relevant policies but improved after segmentation and increasing the number of topics. Table III in Appendix presents a before-and-after visualization of how the model mapped policies/programs to topics in the PIRS scorecard. The document column only contains policies/programs in the PIRS scorecard given to us by the client.

The 15-topic model struggled to confidently assign relevant topics to several policies, often labeling them as "None" with low probability values (e.g., 0.067 across all topics). This indicates that the model was unable to differentiate distinct themes within the dense policy text.

However, the 16-topic model demonstrates a significant improvement in topic classification. For example:

- **Policy 1.1 (Page II-13** was previously unclassified in the 15-topic model but was correctly identified under "Noise Regulations & Land Use" with a **probability of 0.83** in the refined model.
- **Policy 5.1 (Page II-27**, which addresses multi-family density and slope restrictions, initially lacked a meaningful topic assignment. However, the updated model confidently categorized it under "Environmental Conservation & Emergency Response" with a **probability of 0.78**.
- **Policy 1.4 (Page III-28**, relating to city street design, saw an increase in topic clarity, being mapped to "Master Planning & Infrastructure Development" with **0.393 probability** in the improved model.

These results highlight the impact of expanding the topic space and segmenting policies into individual programs. Moving forward, we would like to enhance the efficiency of our analysis because there is still a need for manual verification of the LDA outputs.

D. Processing Time Considerations

Once policies are extracted, the LDA model assigns topics and probability distributions almost instantly. However, if policies are first extracted using an LLM, the total processing time increases due to API rate limits and model response latency. Direct extraction from structured documents is nearly instantaneous, whereas querying an LLM for unstructured policy text may take significantly longer, especially for large datasets.

VI. FUTURE WORK & CONCLUSION

Over the next three months, we plan to continue extracting policies from plans in other counties across California, starting with Napa County. Our rule-based text extraction techniques may require modifications to ensure adaptability across different document formats. Results from LLM queries will need manual error checks by going through each document chunk and marking which policies were read or missed.

From an LDA perspective, we will evaluate how well our topic model generalizes to different counties' planning documents and whether additional refinements, such as adjusting the number of topics or modifying preprocessing steps, are needed. We also aim to automate aspects of LDA topic validation by implementing statistical measures, such as topic coherence and document-topic probability thresholds, to flag uncertain classifications for further review. This will help reduce the need for manual verification while maintaining accuracy in topic assignments.

ACKNOWLEDGMENT

We sincerely thank Bill Siembieda, Margot McDonald, and Andrew Fricker for their guidance and support throughout this project. Their insights into wildfire resilience and policy analysis have been invaluable in shaping our approach.

VII. Appendix

TABLE III
Topic Assignments and Probability Distributions for Policies & Programs Between the 15-Topic and 16-Topic LDA Models

Page	**Policy**	**Topic (15 Model)**	**Probability (15 Model)**	**Topic (16 Model)**	**Probability (16 Model)**
II-13	Policy 1.1: Preserve the rural atmosphere of the community and assure "elbow room" in areas designated for lower density development by guiding new development into the Urban Core to conform to the historic Colony land use patterns of the City and to respect the natural environment, hillside areas, and existing neighborhoods.	None	0.067	Noise Regulations & Land Use	0.83
II-13	[1.1]2: Concentrate higher-density development downtown and within the Urban Core, and focus master-planned commercial uses at distinct nodes along arterial corridors.	None	0.067	Master Planning & Infrastructure Development	0.436
II-27	Policy 5.1: Reduce multi-family densities and increase single-family lot sizes as site slope increases.	None	0.067	Environmental Conservation & Emergency Response	0.78
II-30	[6.4]4: Utilize the Secretary of the Interior's Standards and Guidelines for Rehabilitating Historic Properties to assess proposed improvements to historic properties.	None	0.067	Housing Policy & Neighborhood Compatibility	0.774
III-28	Policy 1.4: Preserve the winding, tree-lined nature of the city street system in hillside areas. Programs: hillsides to preserve rural character and help limit vehicle speed.	None	0.067	Master Planning & Infrastructure Development	0.393