75.06 Organización de datos

Trabajo Práctico 1:
Análisis exploratorio de datos

Grupo: DataMasters

Nombre	Padrón	Mail
Ezequiel Vilardo	104980	ezequielvilardo@gmail.com
Santos Emanuel Amaya	96891	samaya@fi.uba.ar
Rogger Aldair Paredes Tavara	97976	rpardest@fi.uba.ar
Pablo Ariel González	95079	gpabloariel@gmail.com

Github:
https://github.com/EzequielVF/7506-Datos-TP1

1. Introducción.

Se analiza la información relevada de los daños sufridos a nivel infraestructura a causa del terremoto en Nepal del año 2015:

https://www.drivendata.org/competitions/57/nepal-earthquake/data/

Basado en este análisis se quiere prever cómo reducir el impacto en eventos similares.

2. Análisis estructural de los DataSet disponibles.

El análisis en cuestión se realiza utilizando los siguientes archivos:

- train_labels.csv.
- train_values.csv.

2.1 Controles y modificaciones realizadas.

- Se verifica que los dataset no tengan valores faltantes (nulos).
- Se verifica que los dataset no tienen dos valores “building_id” iguales.
- Se reducen los tamaños de ciertos campos enteros, luego de controlar los valores máximos existentes:
    - building_id: int32.
    - geo_level_1_id: int8.
    - geo_level_2_id: int16.
    - geo_level_3_id: int16.
    - count_floors_pre_eq: int8.
    - age: int16.
    - area_percentage: int8.
    - height_percentage: int8.
    - count_families: int8.
    - has_secondary_use: int8.
    - has_secondary_use_agriculture: int8.
    - has_secondary_use_hotel: int8.
    - has_secondary_use_school: int8.
    - has_secondary_use_industry: int8.
    - has_secondary_use_health_post: int8.
    - has_secondary_use_gov_office: int8.
    - has_secondary_use_use_police: int8.
    - has_secondary_use_other: int8.
    - has_secondary_use_rental: int8.
    - has_secondary_use_institution: int8.
    - has_superstructure_adobe_mud: int8.
    - has_superstructure_mud_mortar_stone: int8.

- has_superstructure_stone_flag: int8.
- has_superstructure_cement_mortar_stone: int8.
- has_superstructure_timber: int8.
- has_superstructure_bamboo: int8.
- has_superstructure_rc_non_engineered: int8.
- has_superstructure_rc_engineered: int8.
- has_superstructure_other: int8.
- has_superstructure_mud_mortar_brick: int8.
- has_superstructure_cement_mortar_brick: int8.

- Se definieron todos los campos posibles del tipo Object como Category con el fin de mejorar el rendimiento:

- land_surface_condition.
- foundation_type.
- roof_type.
- ground_floor_type.
- other_floor_type.
- position.
- plan_configuration.
- legal_ownership_status.

2.2 Beneficio obtenido.

El tamaño del dataset se redujo de 77,5 MB aproximadamente. a sólo 11,2 MB aproximadamente. Esto equivale a una reducción del 85% en el tamaño del dataset.

3. Análisis exploratorio.

Para lograr un análisis detallado, separamos la información en distintos bloques de datos.

3.1 Epicentro.

El objetivo de este bloque, de ser posible, es determinar el epicentro del terremoto.

3.1.1 Hipótesis - Preguntas.

¿Dónde fue el epicentro del terremoto? ¿Cuáles fueron las áreas más afectadas?

3.1.2 Desarrollo.

Se analizaron las columnas asociadas a los 3 niveles de geolocalización. Como no se logró encontrar una relación estrecha entre niveles, se optó por calcular el epicentro a partir del geo level 1.

Se inició con el análisis de la ubicación contra los daños sufridos en cada nivel conformando el gráfico de la Figura 1.

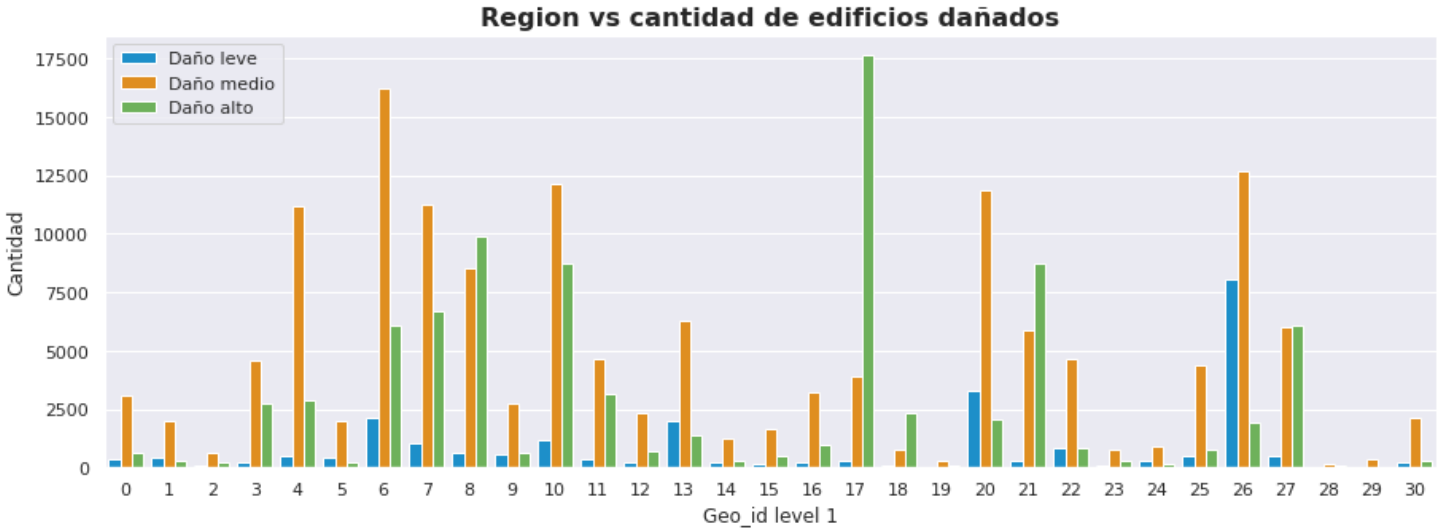

Figura 1: Cantidad de edificios dañados por nivel y localización.

Analizando el daño en detalle, se armó un gráfico (Figura 2) con el fin de comparar los valores normalizados. Con estos gráficos ya se puede visualizar cuáles áreas fueron las más afectadas.

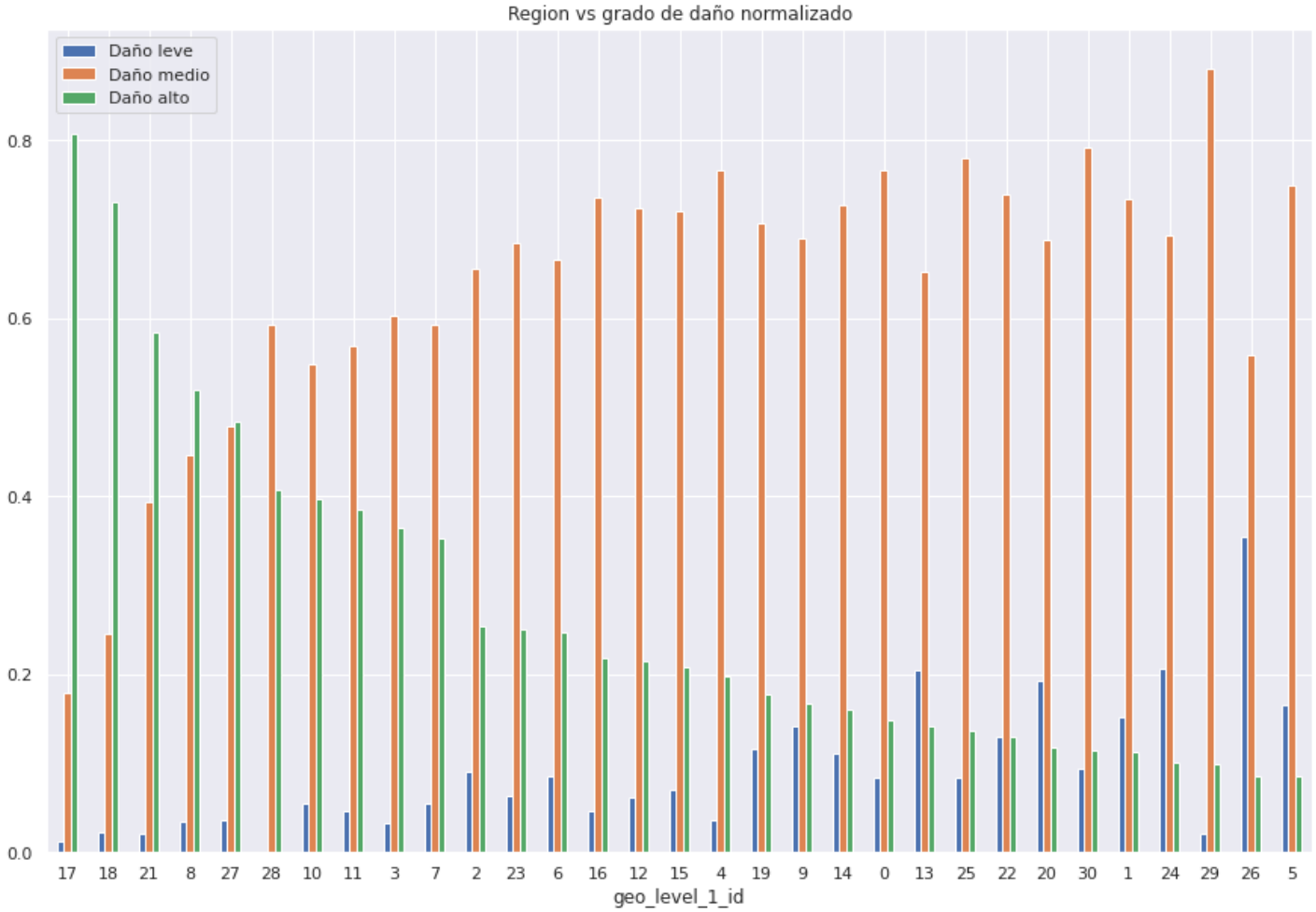

Figura 2: Cantidad de edificios dañados por nivel y localización, valor normalizado.

Con la información obtenida, se construye el gráfico número 3 del tipo radar para terminar de ilustrar cuáles áreas fueron las más afectadas:

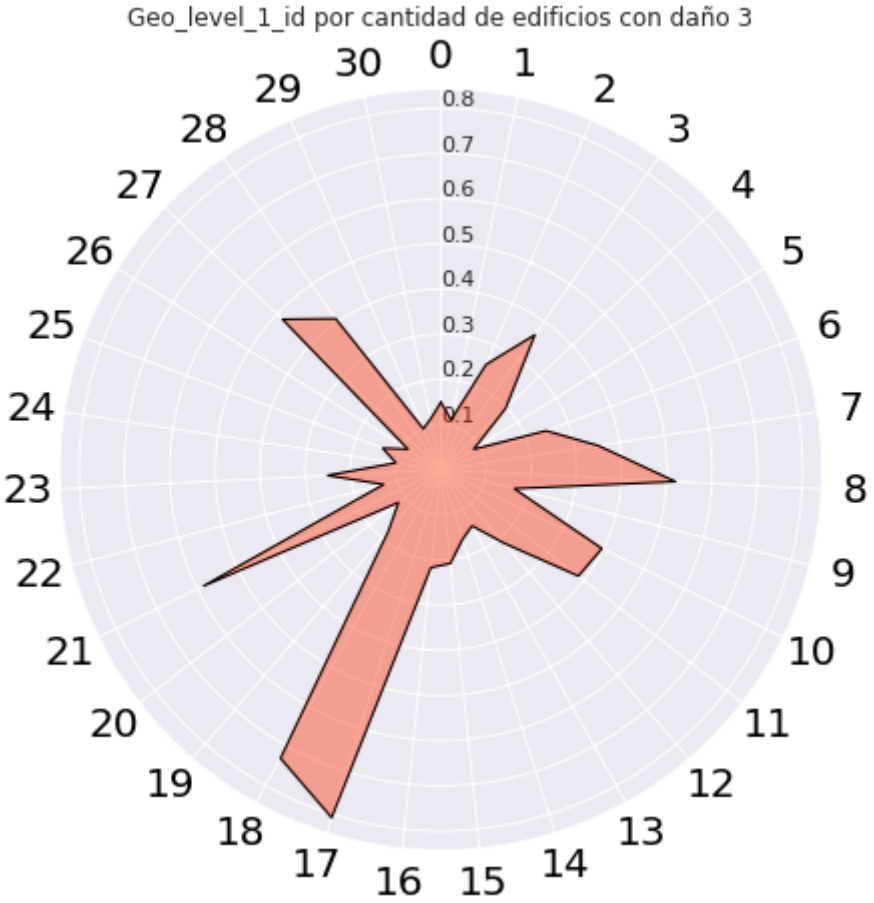

Figura 3: Cantidad de edificios por daño 3 en cada localización.

3.1.3 Conclusión.

Según lo desarrollado se detecta que las áreas de nivel 1 denominadas 17 y 18 fueron las más afectadas, entendiendolas como el epicentro del terremoto.
Luego, según el daño recibido, se interpreta que existen dos anillos alrededor conformados por las áreas 21, 27 y 28 como un primer anillo y las áreas 3, 8, 10 y 11 como segundo anillo. Luego se encontrarían el restos de las áreas disponibles.

3.2 Impacto gubernamental (infraestructura).

En este bloque se busca determinar qué sector se vio más afectado y por lo cual va a necesitar un mayor apoyo por parte del gobierno.

3.2.1 Hipótesis - Preguntas.

¿Cuál o cuáles fueron los sectores más afectados? ¿En qué nivel?

3.2.2 Desarrollo.

Se obtiene el uso de cada edificio y se verifica cuantos fueron afectados y que nivel de daño sufrieron (Figura 4).

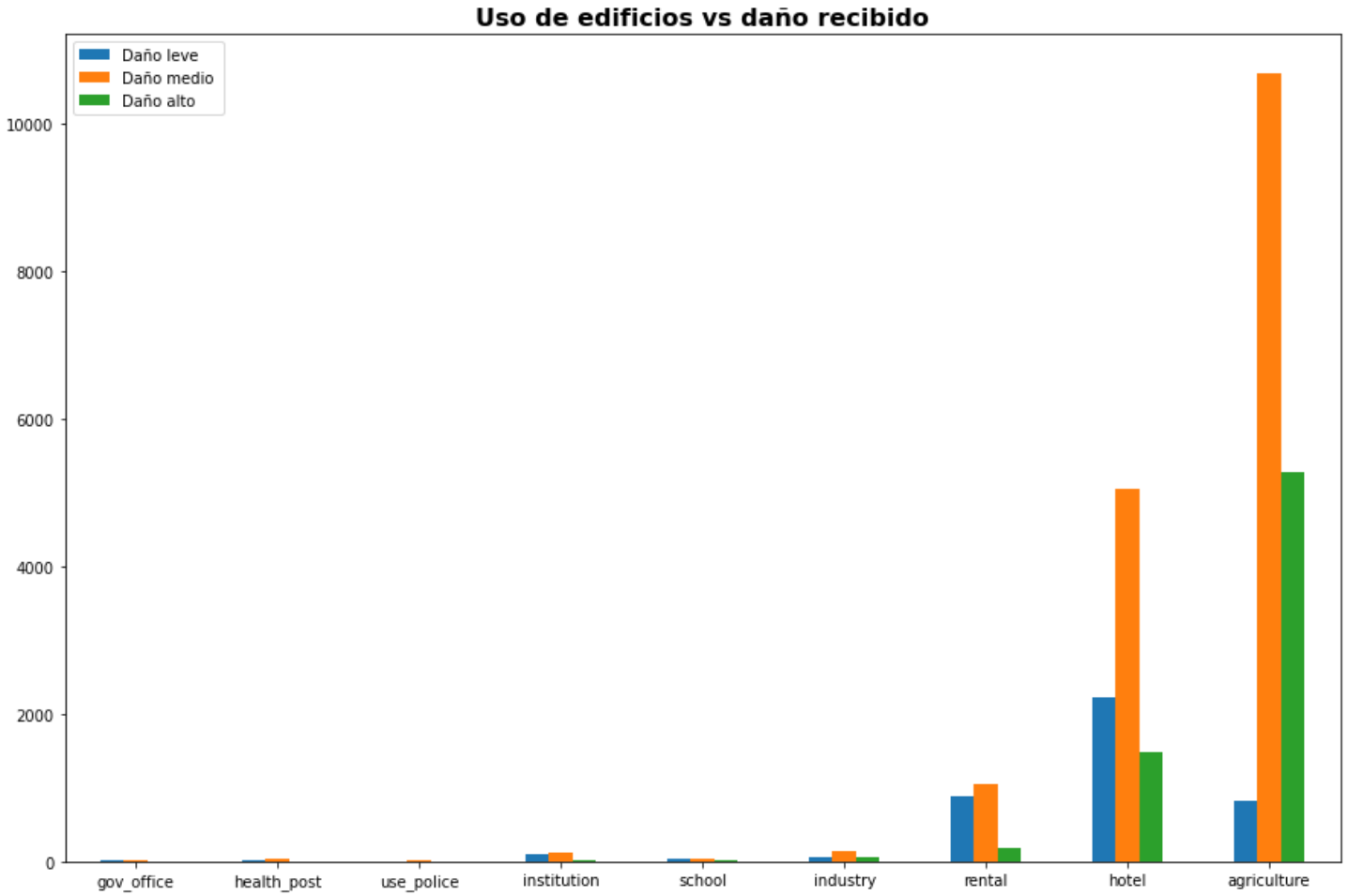

Figura 4: Daño recibido en cada edificio según uso.

Luego, se calcula y compara las cantidades normalizadas en el gráfico de la Figura 5 con cada uso de los edificios:

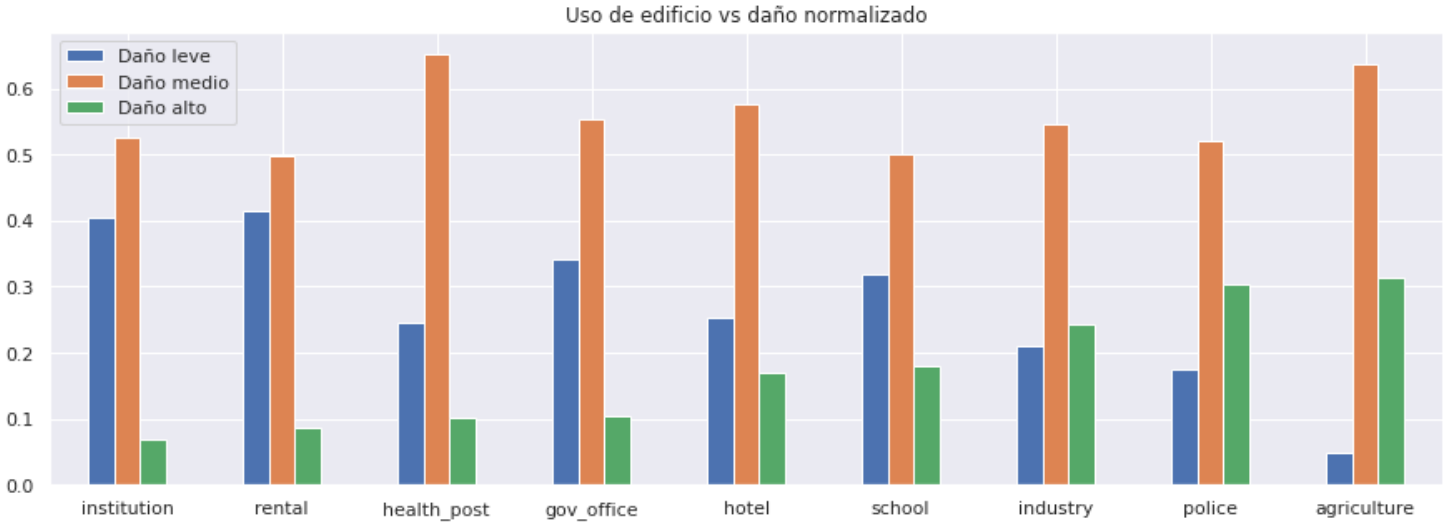

Figura 5: Daño normalizado recibido en cada edificio según uso.

3.2.3 Conclusión.

Claramente se verifica que el sector de agricultura es el más afectado con un total de:

- Daño Bajo: 829 edificios.
- Daño Medio: 10679 edificios.
- Daño Alto: 5269 edificios.

En una segunda instancia se encuentran los edificios del tipo Hotel, cuyas cantidades son las siguientes:

- Daño Bajo: 2216 edificios.
- Daño Medio: 5058 edificios.
- Daño Alto: 1489 edificios.

El resto de los edificios, si bien fueron afectados, cada uso representa menos del 10% en comparación con las 2 categorías ya nombradas.
Por este motivo se entiende que dichas actividades son las principales a tener en cuenta en un plan de reconstrucción.

3.3 Construcción.

Dentro de este bloque se analizaron los datos asociados al tipo de estructura, cimientos, tipos de techo y dimensiones.

3.3.1 Hipótesis - Preguntas.

¿Cúal o cuáles fueron las peores construcciones, es decir las más afectadas? ¿Cuáles fueron las menos afectadas? ¿Existe una estructura/construcción recomendada?

3.3.2 Desarrollo.

3.3.2.1 Impacto factor antigüedad.

A partir de los datos obtenidos se diseña el gráfico que se visualiza en la Figura 6 y se verifica en los últimos años una reducción en el nivel de daño recibido, esto probablemente se deba a una mejoría en la tecnología de la construcción

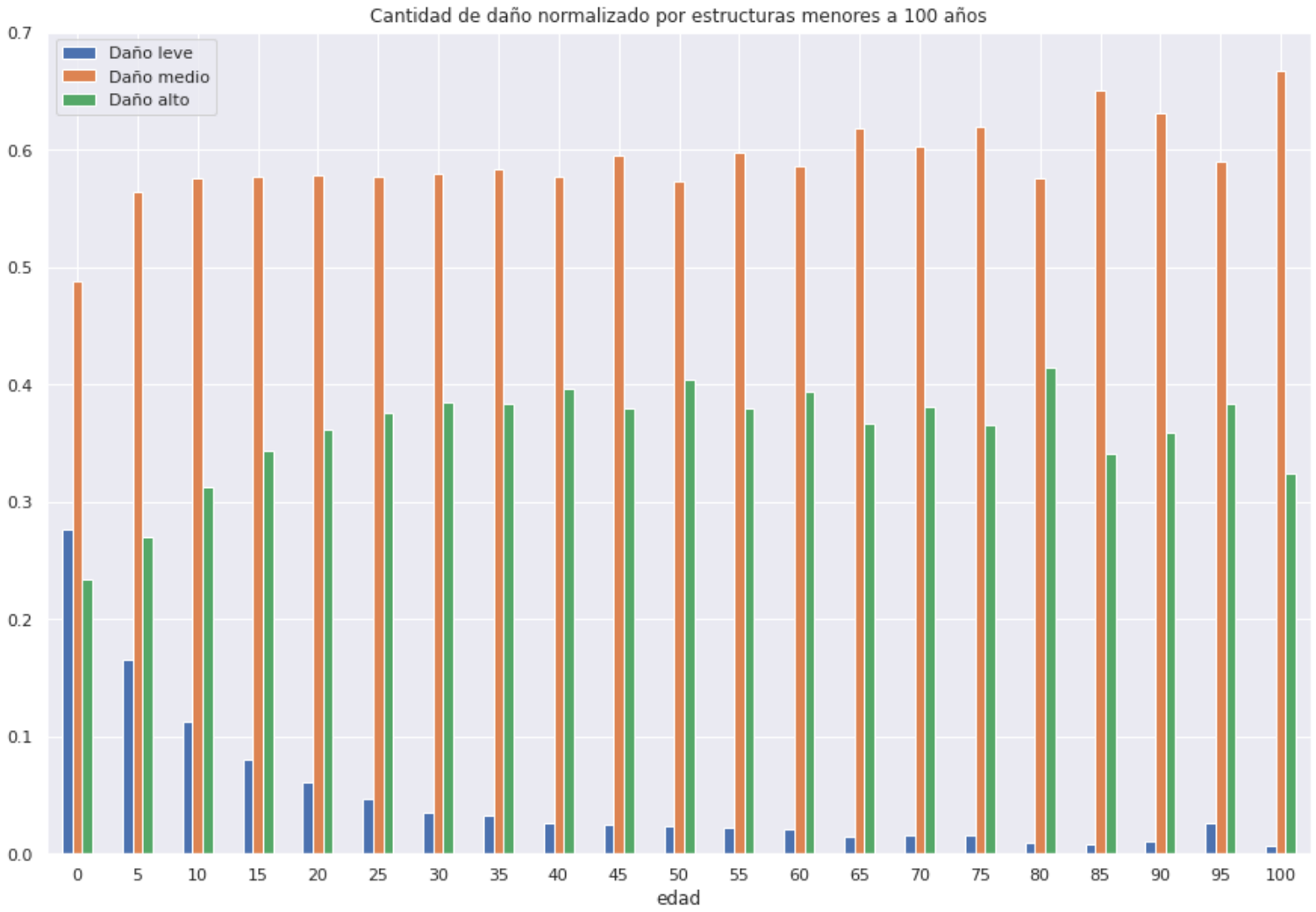

Figura 6: Daño normalizado recibido en cada edificio por antigüedad menor o igual a 100 años.

Por otro lado, el análisis de daño sufrido para edificios de mayor antigüedad de la Figura 7 no define ningún patrón particular y entendemos que esto se dá por la escasa cantidad de edificios de este tipo.

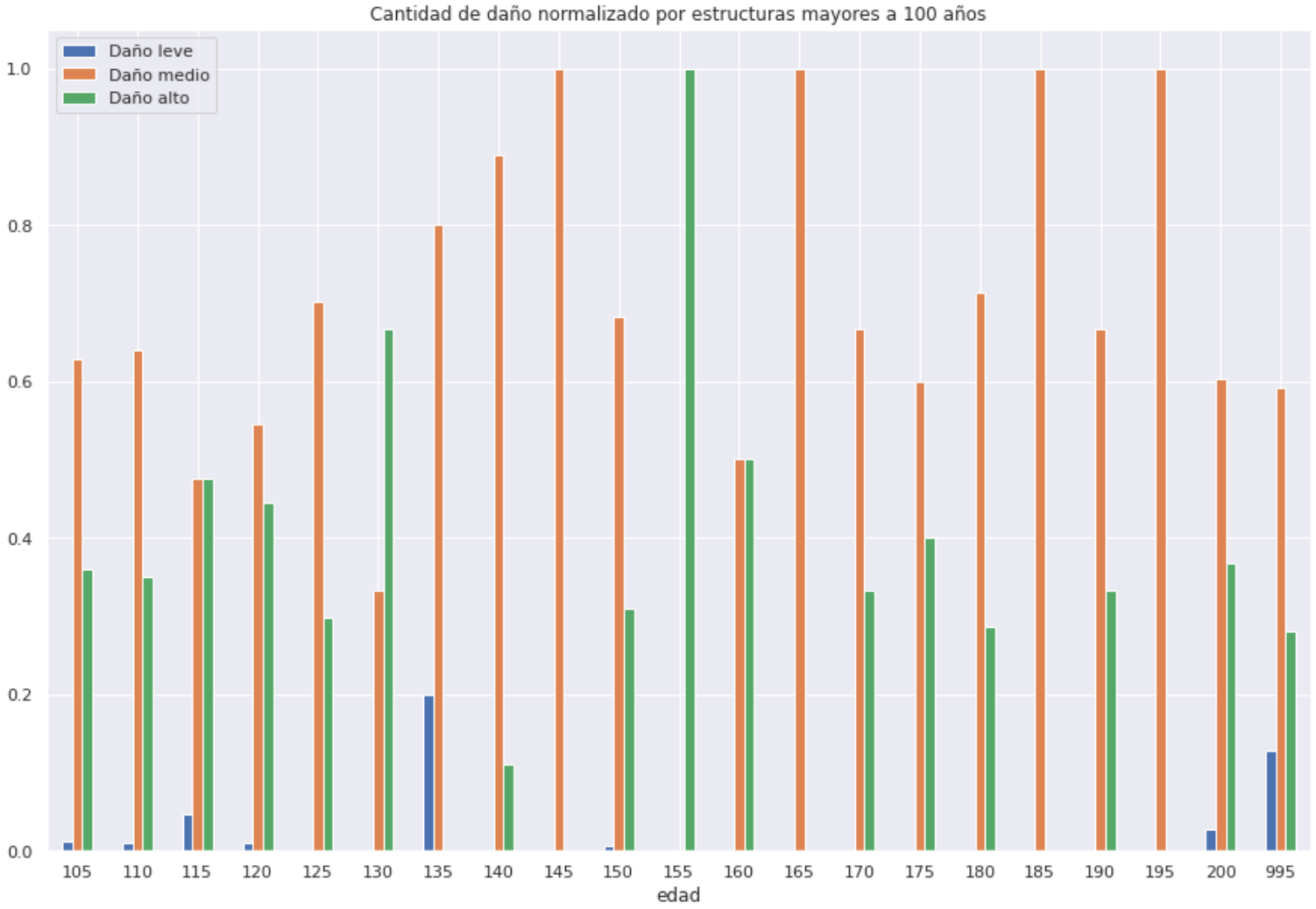

Figura 7: Daño normalizado recibido en cada edificio por antigüedad mayor a 100 años.

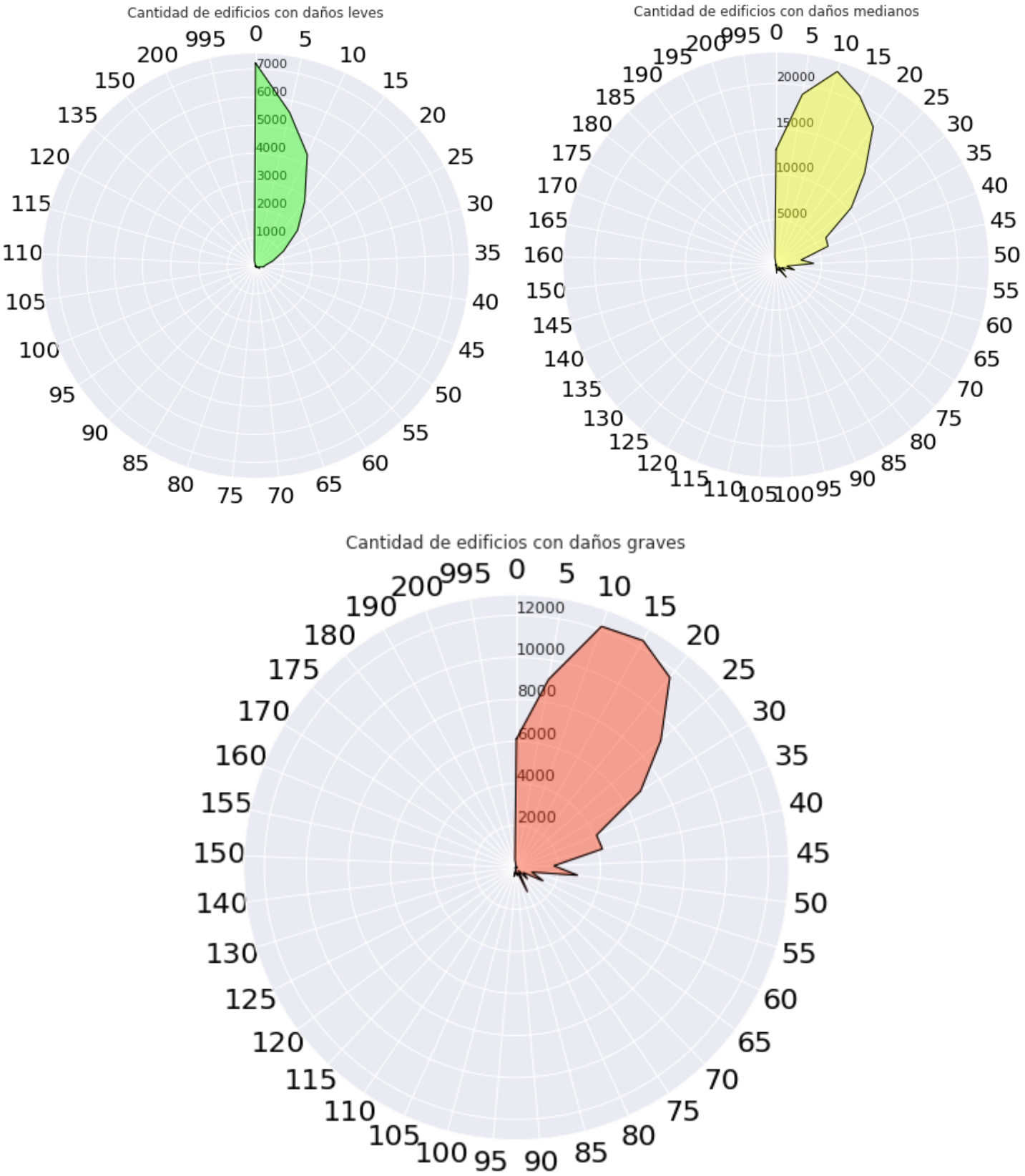

Figura 8: Cantidad de edificios por nivel de daño y antigüedad.

En la Figura 8 se visualiza claramente una mejora asociada a la innovación en edificios nuevos sumado al hecho de no tener prácticamente desgaste en los materiales por el tiempo transcurrido.

3.3.2.2 Impacto factor cantidad de pisos y altura.

Se busca entender si existe una correlación entre la cantidad de pisos o la distribución de pisos en altura junto con el daño recibido.
Para entender esta posibilidad, se realizó una análisis nominal (Figura 9) y un análisis normalizado (Figura 10) del daño recibido según la cantidad de pisos del edificio en cuestión.
Si bien se descartaron los valores de 8 y 9 piso, ya que existía sólo uno de cada uno, el resto de los valores no devolvieron ningún indicador de que efectivamente exista una correlación.

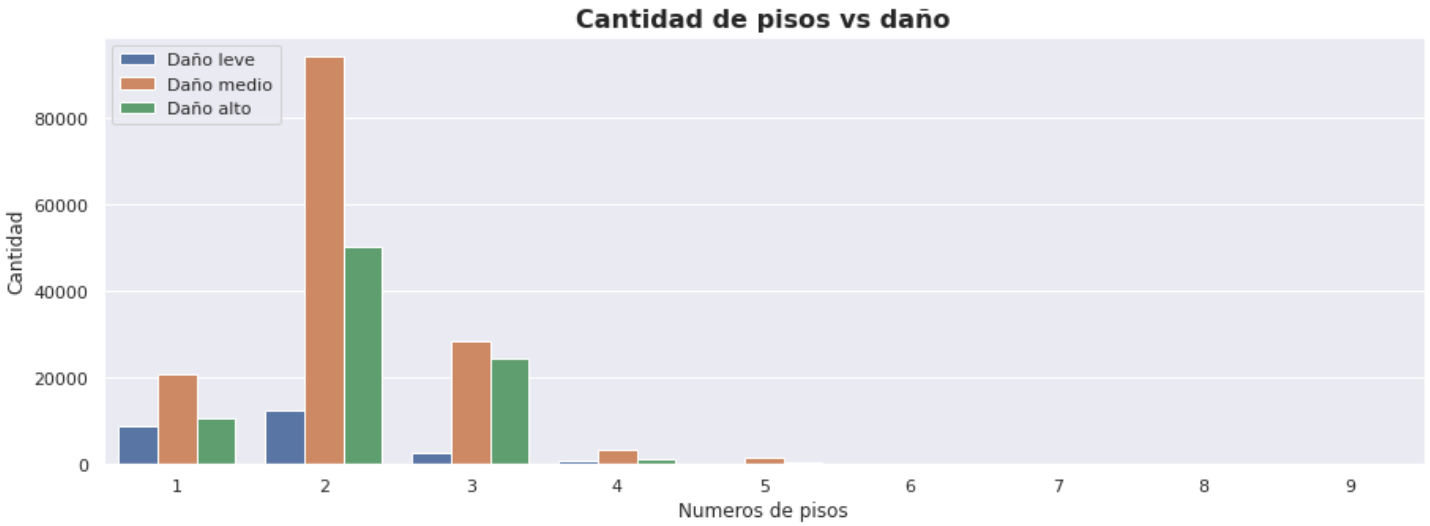

Figura 9: Daño sufrido en cada edificio por cantidad de pisos.

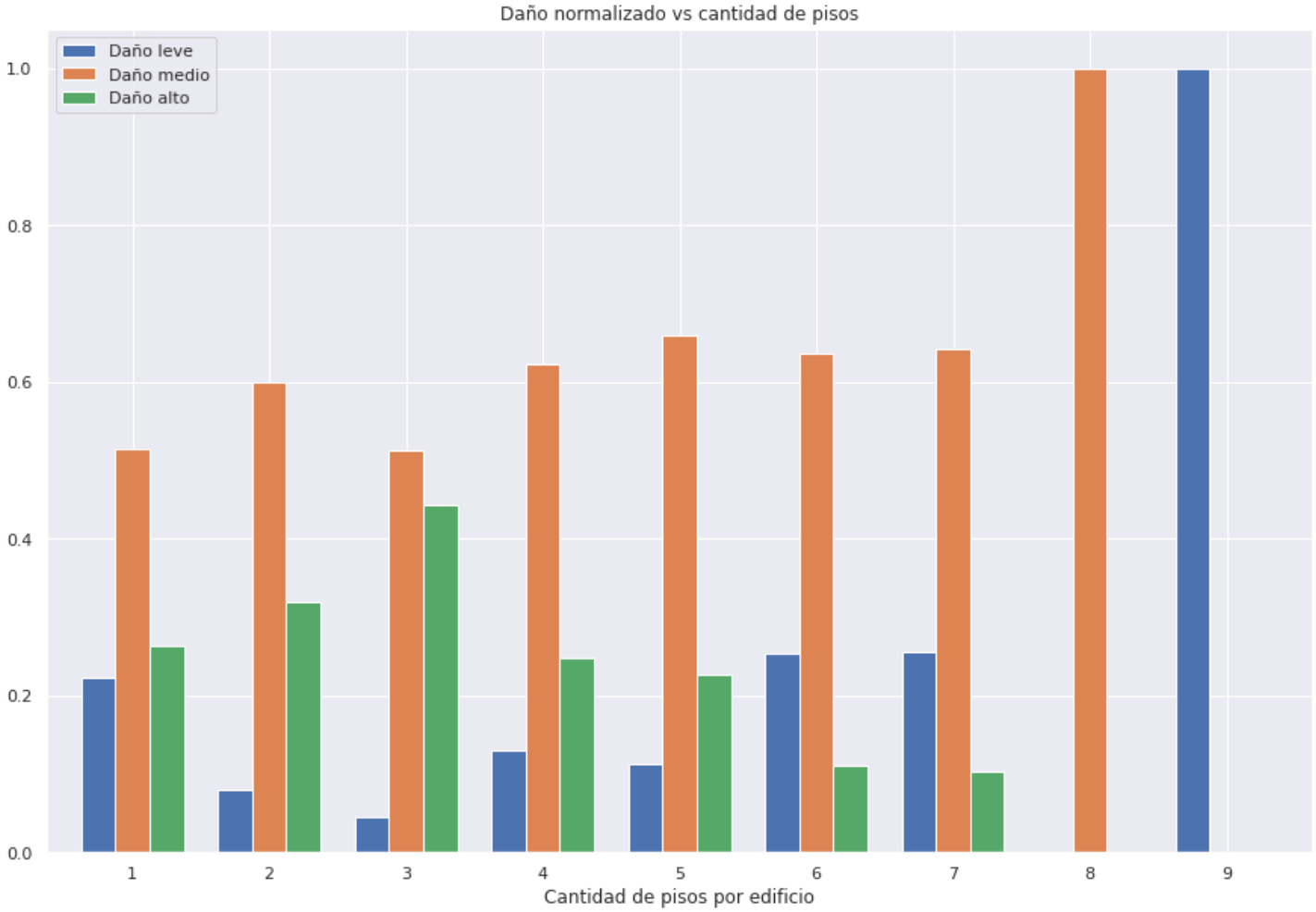

Figura 10: Daño sufrido en cada edificio por cantidad de pisos.

Por otro lado, en relación a la distribución de pisos y altura (se calcula el promedio de ambos atributos para cada edificio) se encuentra una relación de disminución cuando los valores se acercan a la franja de 2.5 - 2.6, lo cual se visualiza con el gráfico 11.
Sin embargo esta relación no es lo suficientemente fuerte para marcar una tendencia.

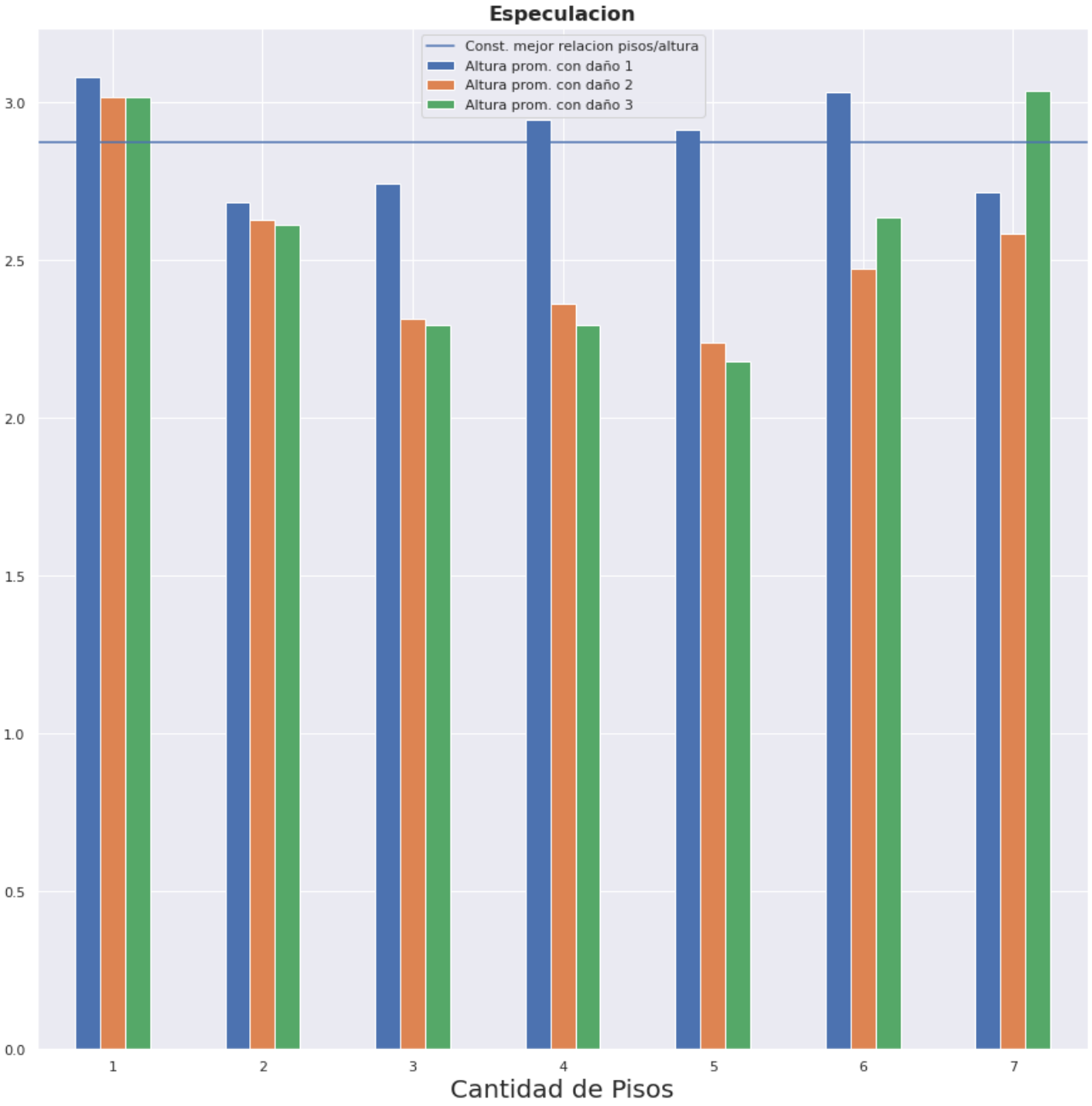

Figura 11: Cálculo de una posible altura óptima para los pisos.

3.3.2.3 Impacto material estructura.

Se analizó la correlación entre el material de la estructura del edificio y el daño recibido en los gráficos de las Figuras 12 y 13. Esta relación es fundamental ya que demuestra cómo el material sufre o no las consecuencias.

Para este caso no solo se analizó el daño en relación al material, sino que además se le aplicó la información obtenida previamente para determinar el epicentro y verificar si existían materiales más resistentes dentro o fuera del epicentro y si convenía algún material en particular según el área donde se disponga el edificio.

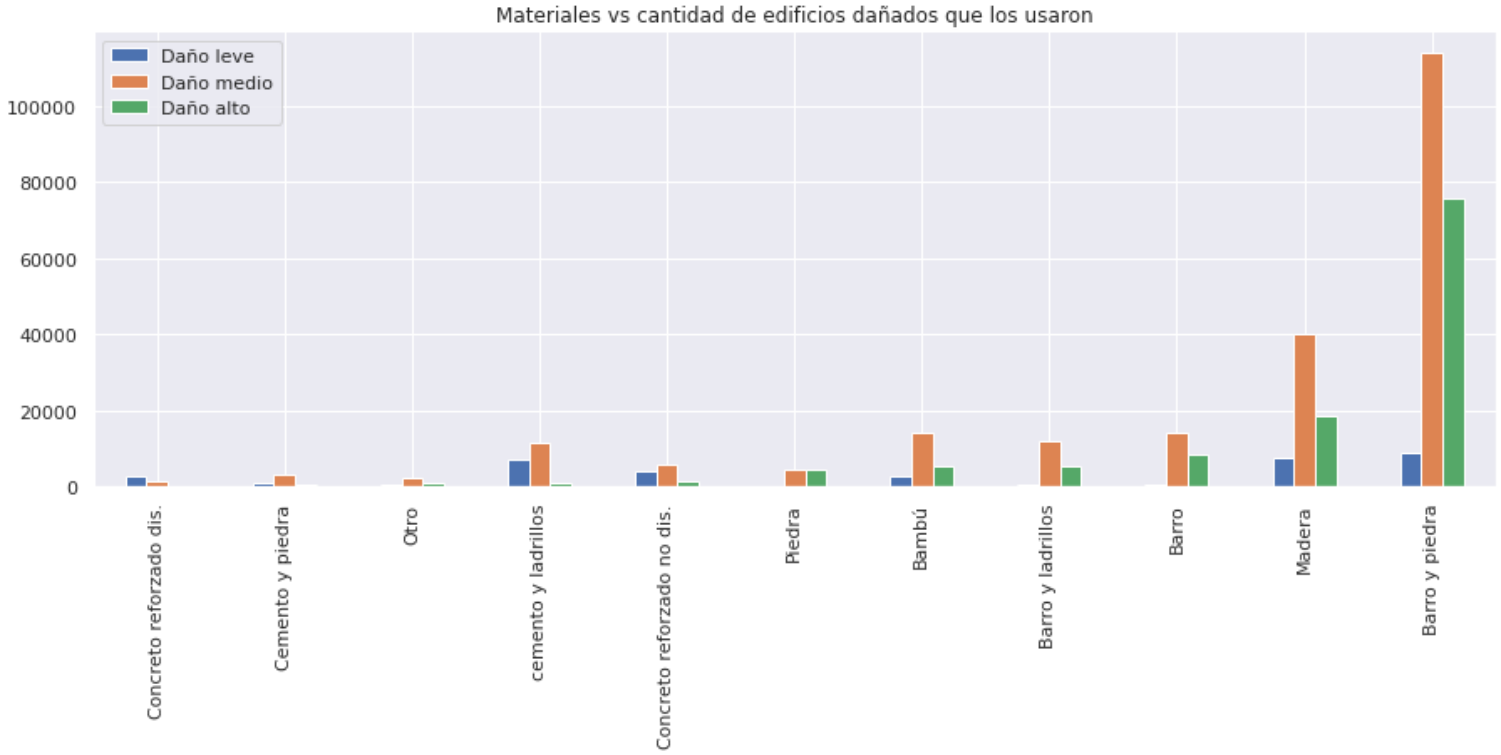

Figura 12: Cantidad de edificios por nivel de daño sufrido y material de estructura..

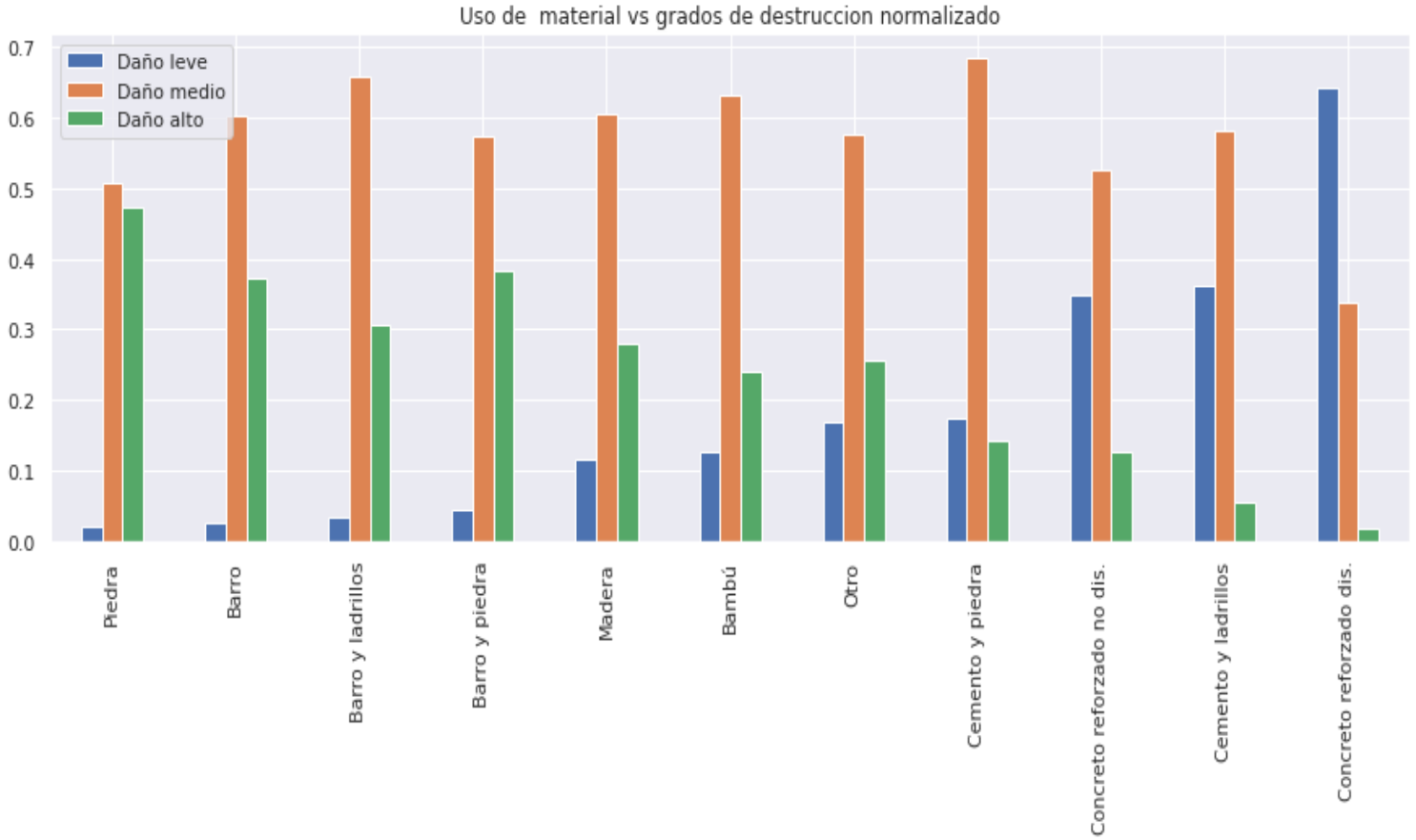

Figura 13: Edificios por nivel de daño sufrido normalizado y material de estructura.

Aplicando la estructura del epicentro obtenido, se define el gráfico de la Figura 14.

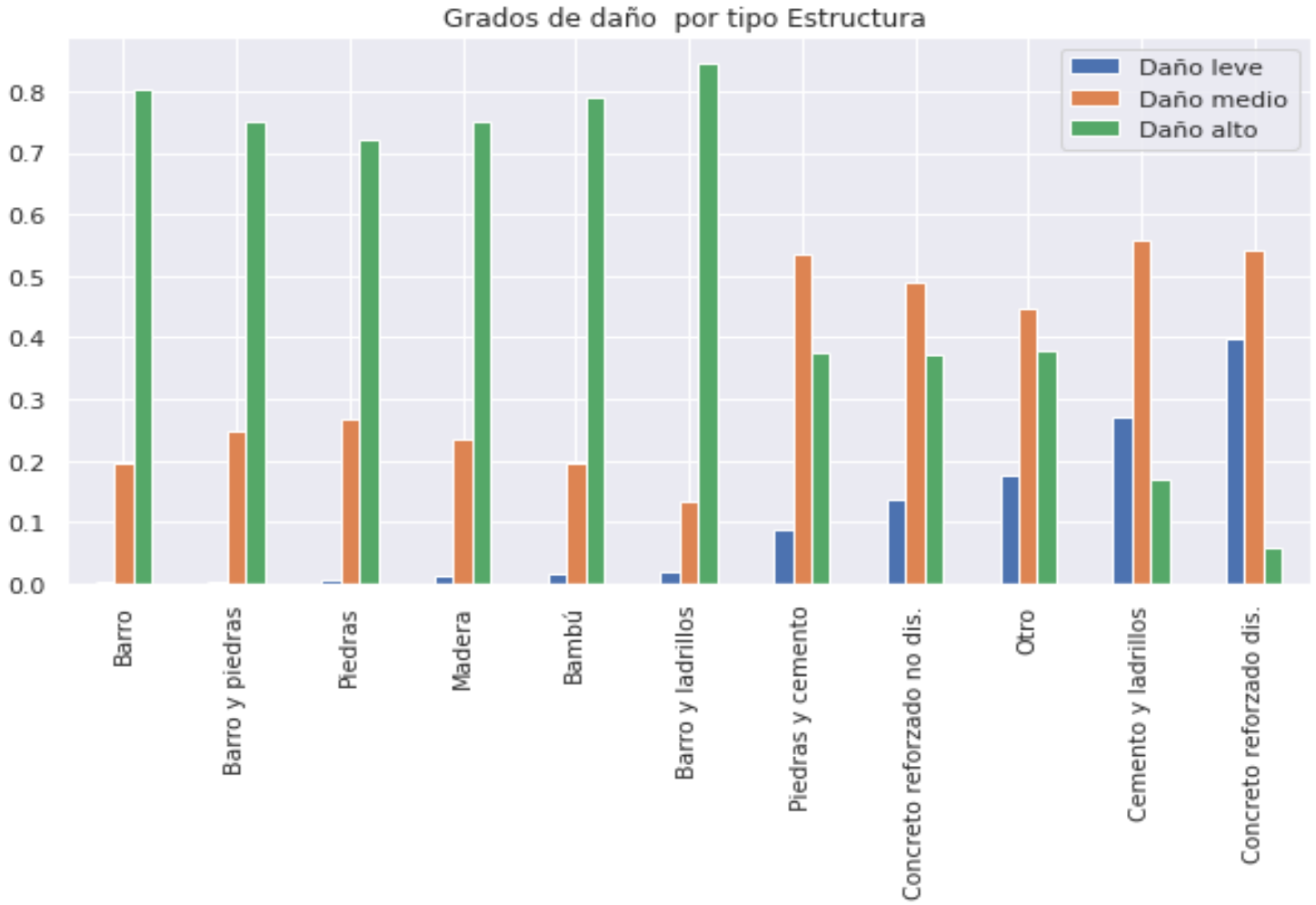

Figura 14: Cantidad de edificios por nivel de daño sufrido y material de estructura dentro del epicentro.

Si bien este análisis cruzado dió algunos valores diferentes, a fines prácticos se visualizaron los mismos resultados. El mejor material para la estructura resulta ser el Concreto Reforzado Diseñado y la construcción de peores resultados fueron la de Barro y la de Barro y Piedra.

3.3.2.4 Impacto material techo.

En relación a los datos acerca de los diferentes tipos de techos se aplicó la misma lógica de análisis de las estructuras. Se revisó la información nominal (Figuras 15 y 16), normalizada (Figura 17 y 18) y se aplicó la lógica de epicentro (Figura 19 y 20)

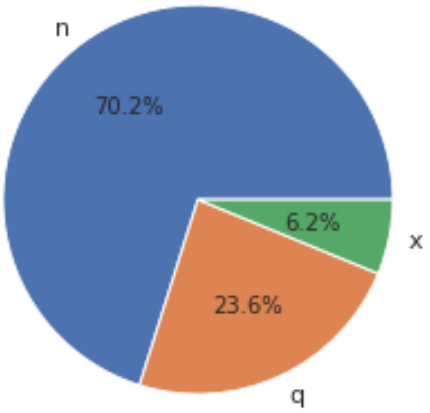

Figura 15: Proporcionalidad de edificios dañados por tipo de material del techo

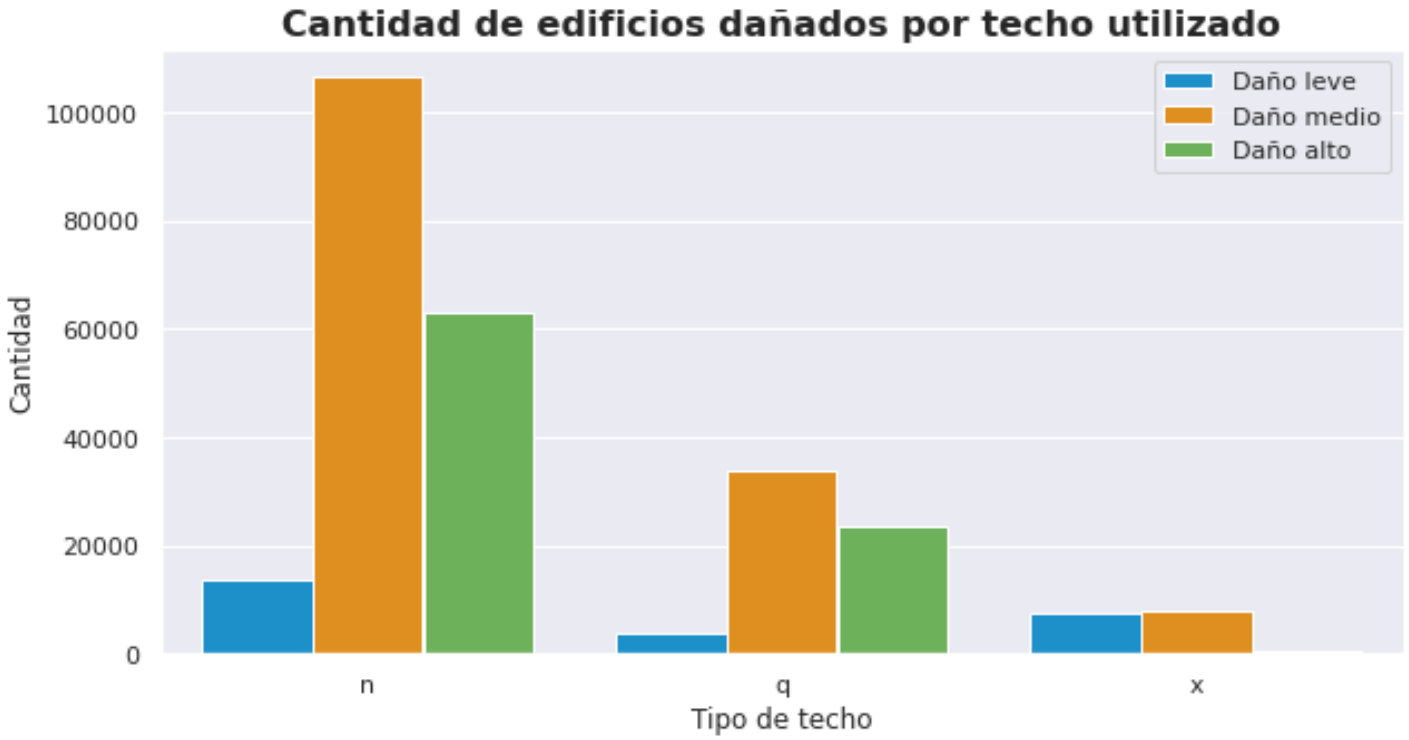

Figura 16: Cantidad de edificios dañados por nivel y tipo de material del techo

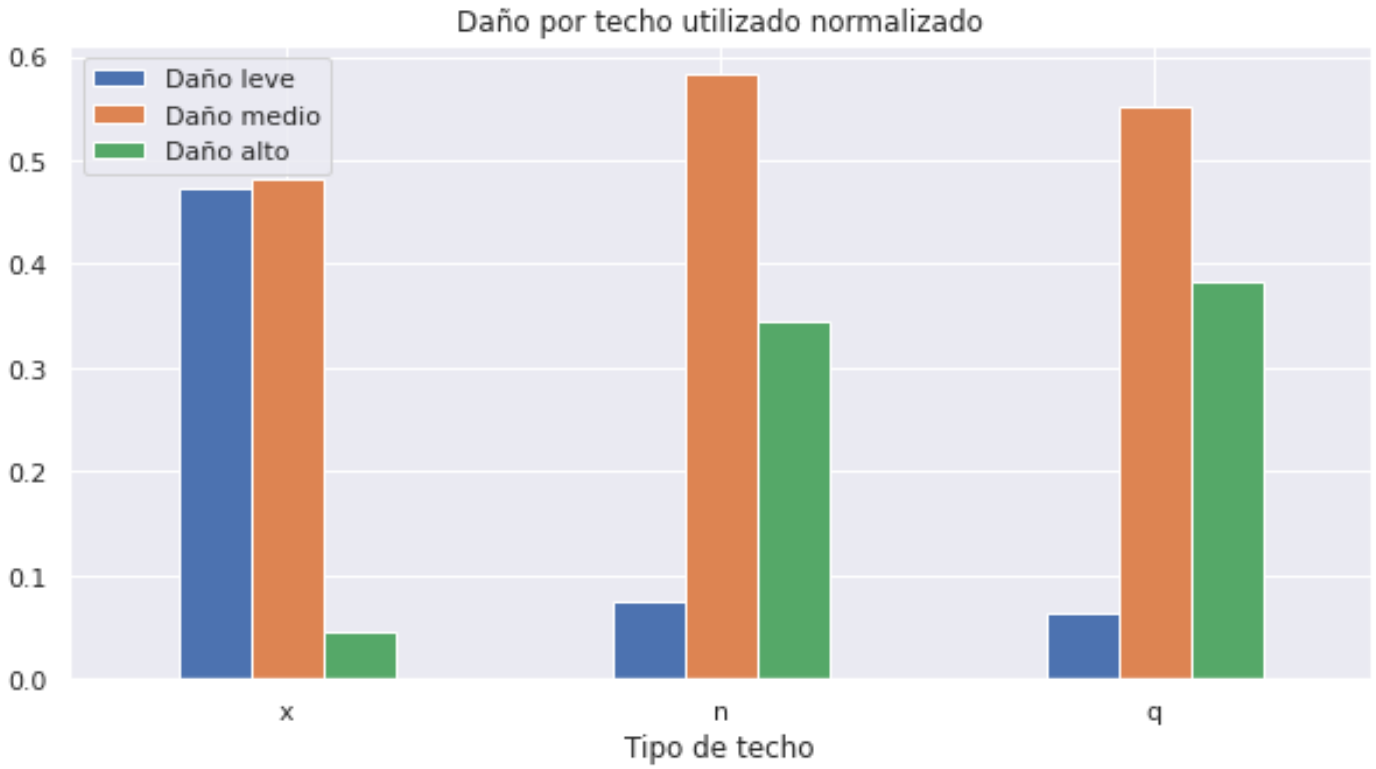

Figura 17: Cantidad de edificios dañados por nivel y tipo de material del techo, valor normalizado.

Se verifica cual es el impacto normalizado del daño por cada tipo de techo y por ubicación geográfica con el objetivo de verificar su correlación.

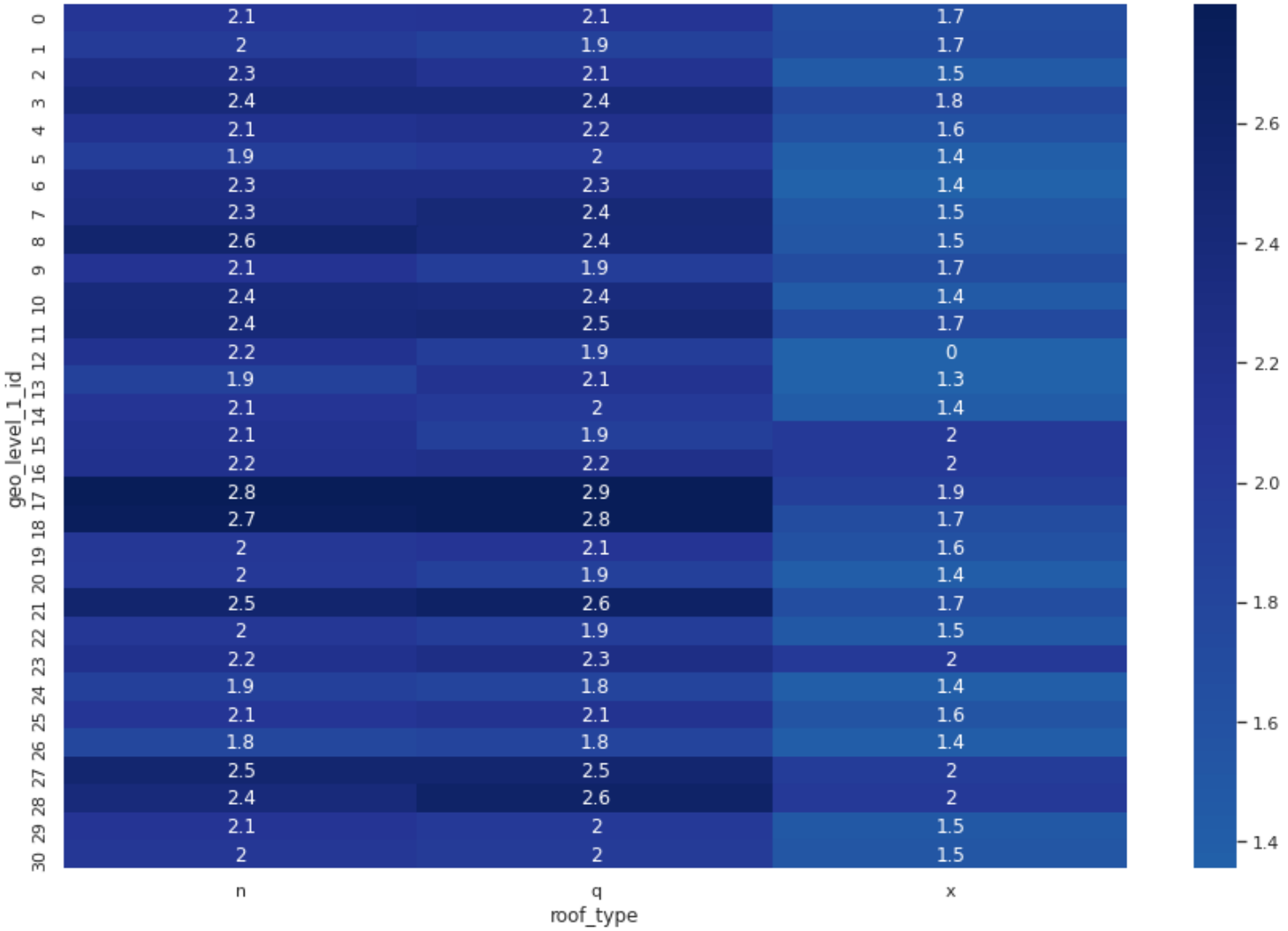

Figura 18: Daños sufridos normalizados por ubicación y tipo de techo.

Se toman los datos del epicentro previamente definidos para aplicarlos como filtros en los gráficos ya analizados con el fin de detectar diferencias en relación a este punto.

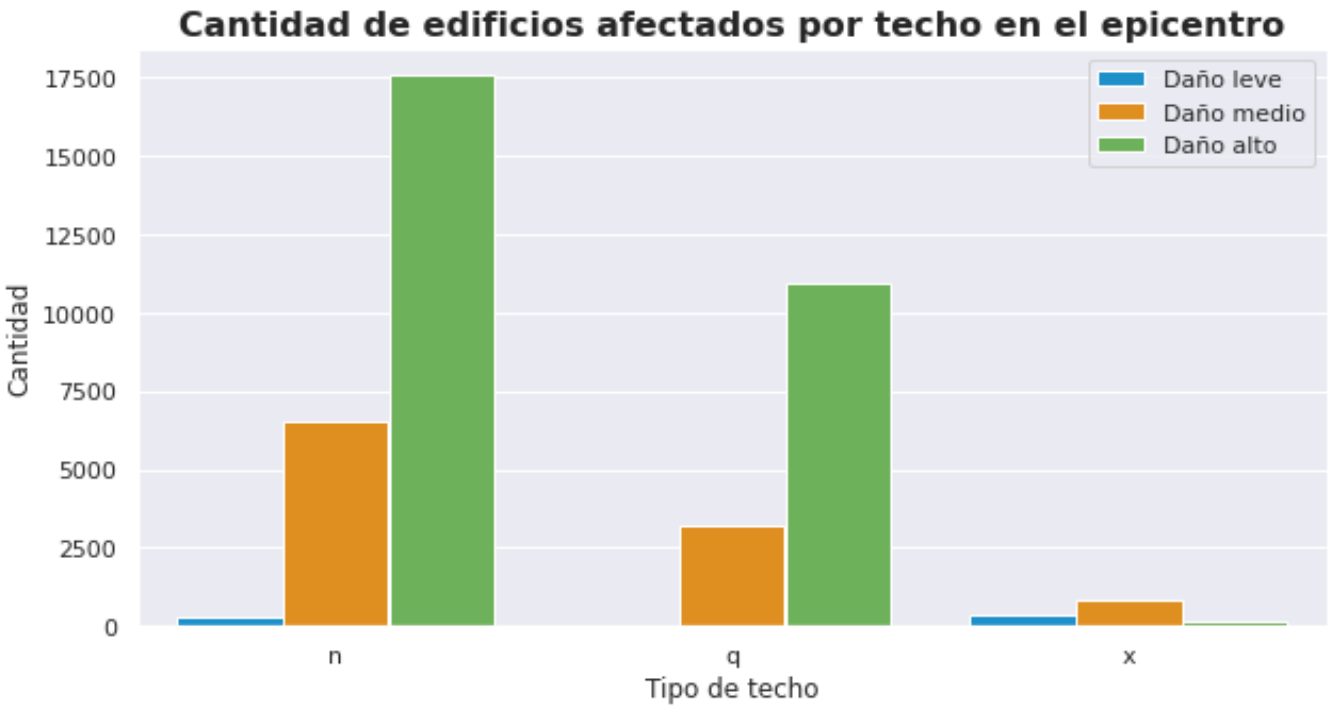

Figura 19: Cantidad de edificios dañados por nivel y tipo de techo dentro del epicentro.

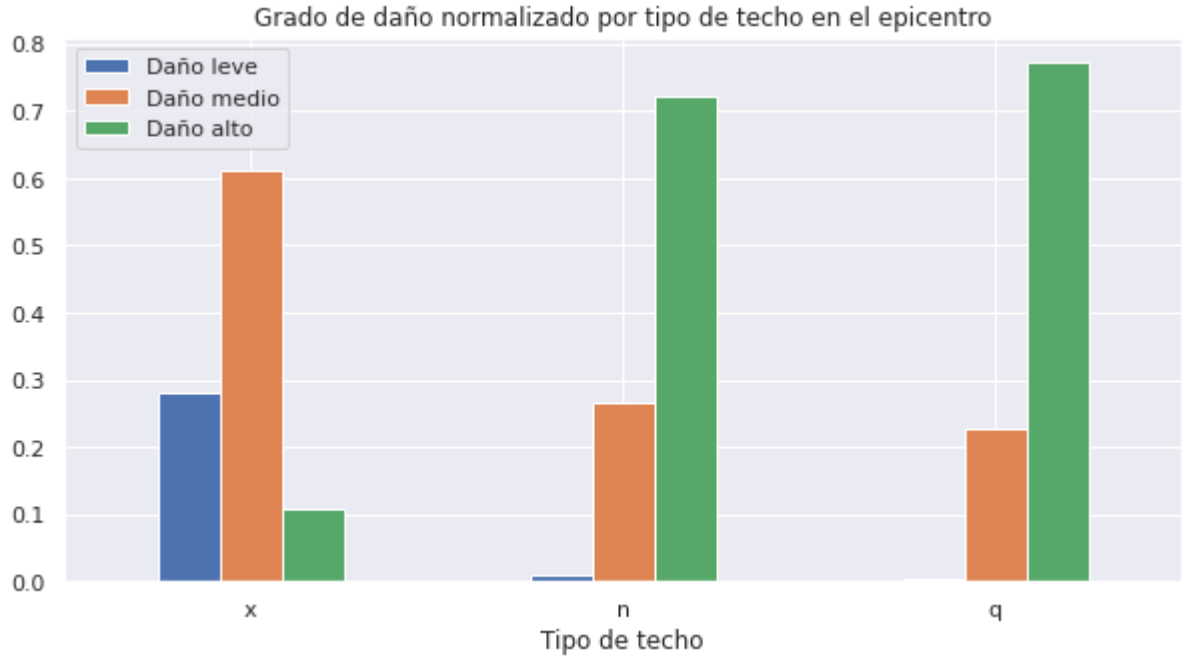

Figura 20: Cantidad normalizada de edificios dañados por nivel y tipo de techo dentro del epicentro

Ya sea dentro o fuera del epicentro el tipo de techo denominado “x” demostró ser el de mejor resultado.

3.3.2.5 Impacto material cimientos.

Se aplicó la misma lógica ya utilizada para determinar el rendimiento de cada clase de cimientos utilizadas tanto dentro como fuera del epicentro.

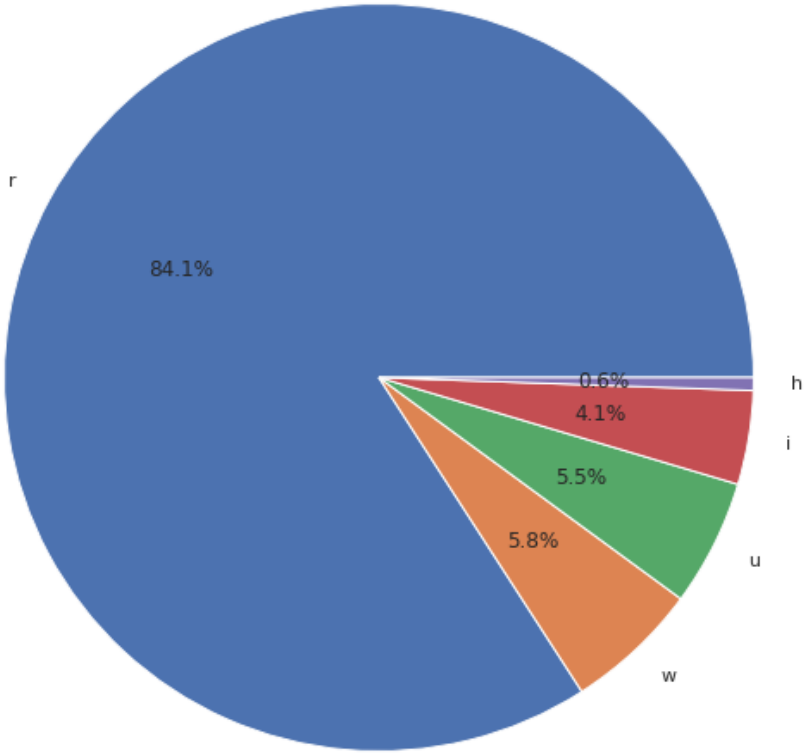

Figura 21: Proporcionalidad normalizada de edificios dañados por material de cimientos.

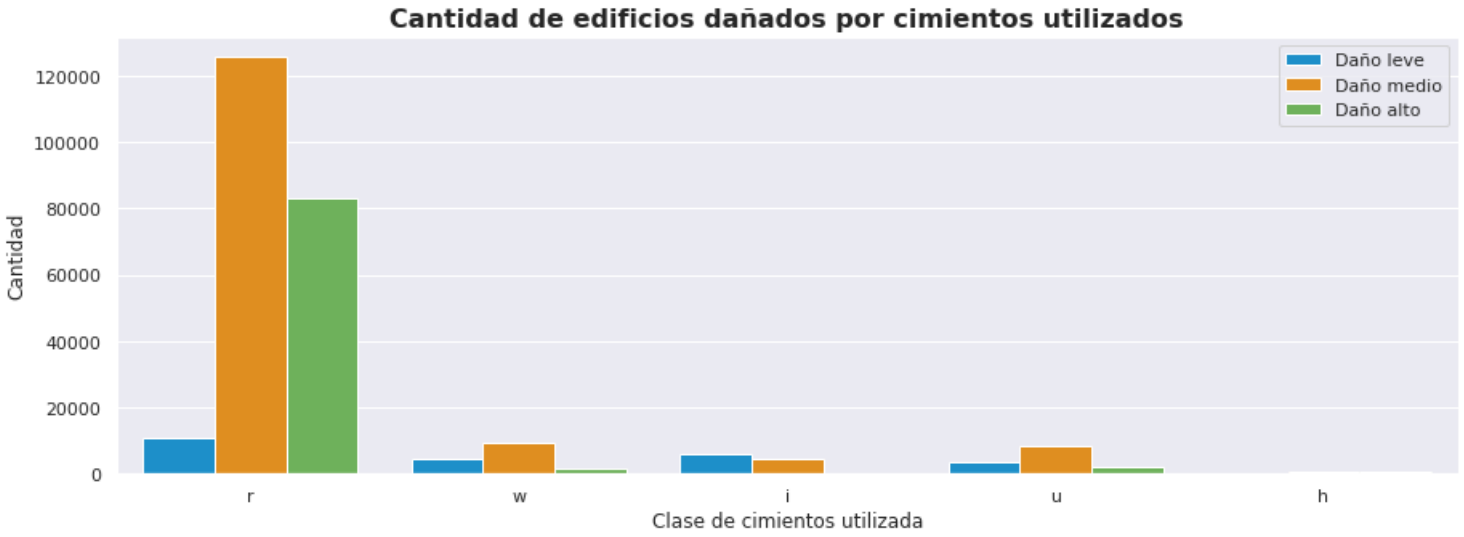

Figura 22: Cantidad de edificios dañados por nivel y material de cimientos.

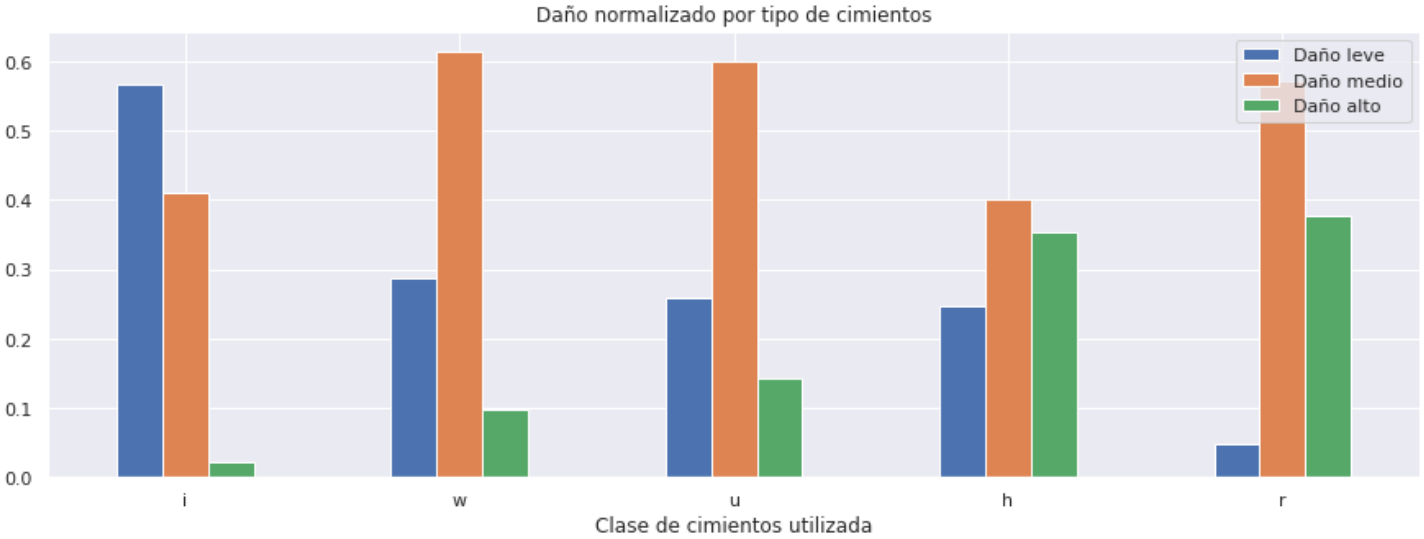

Figura 23: Cantidad de edificios dañados normalizados por nivel y material de cimientos.

Figura 24: Cantidad de edificios dañados por nivel y material de cimientos en epicentro.

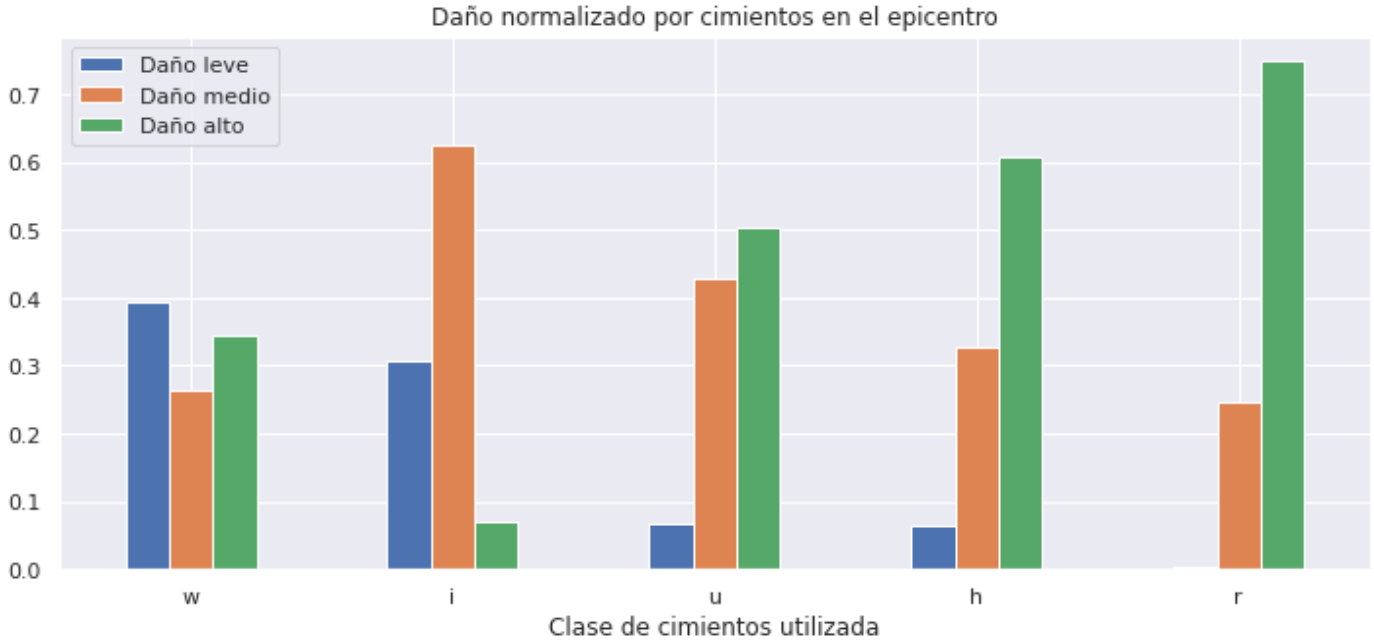

Figura 25: Cantidad de edificios dañados normalizados por nivel y material de cimientos en el epicentro.

Si bien hay valores que poseen muy poco daño a nivel cantidad, entendemos que la clase de cimientos que tuvo el mejor resultado son las de tipo “i” ya que en comparación con el resto prácticamente no sufrió daños del nivel más alto.

3.3.2.6 Impacto formato de construcción.

De manera análoga a los puntos anteriores, se ejecutó el análisis del formato de construcción.
En este indicador los valores normalizados (Figura 27 y 29) tomaron una mayor preponderancia en relación a los valores nominales (Figura 26 y 28) ya que todos los valores distintos de “d” tenían cantidades similares y muy inferiores a la cantidad de edificios dañados con dicho formato. Lo cual complejiza analizarlo a través de un gráfico nominal.

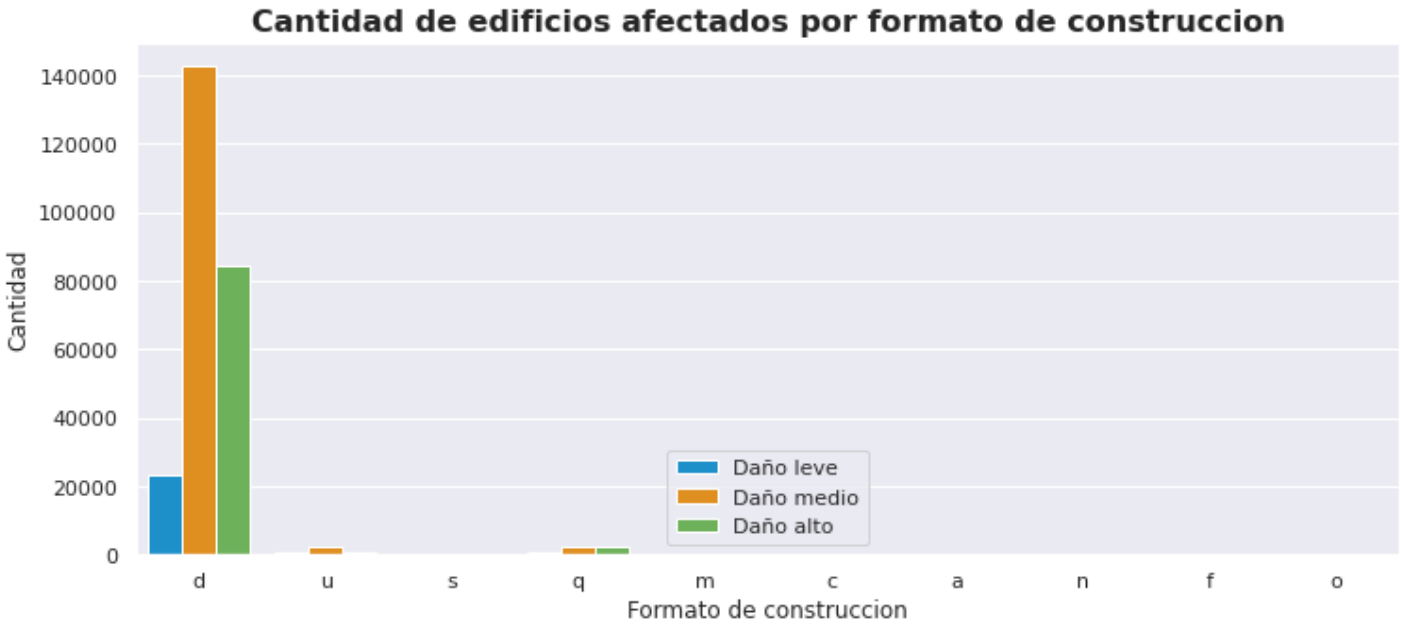

Figura 26: Cantidad de edificios dañados por nivel y formato de construcción.

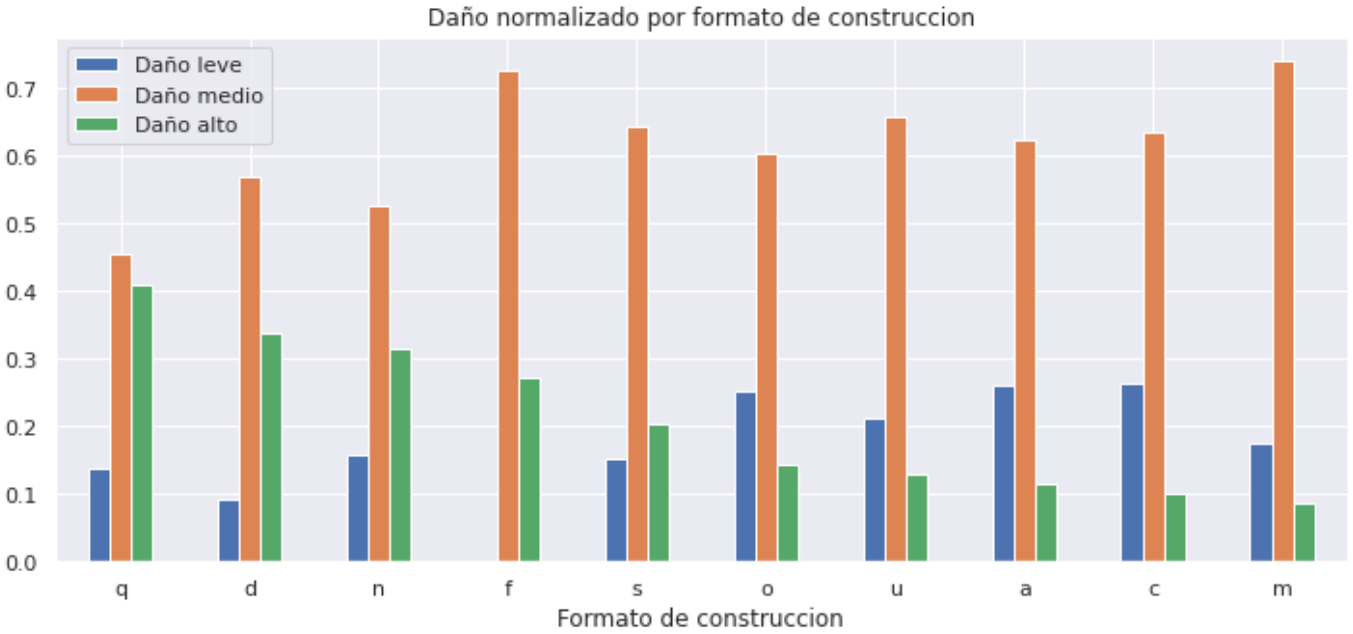

Figura 27: Cantidad de edificios dañados normalizados por nivel y formato de construcción.

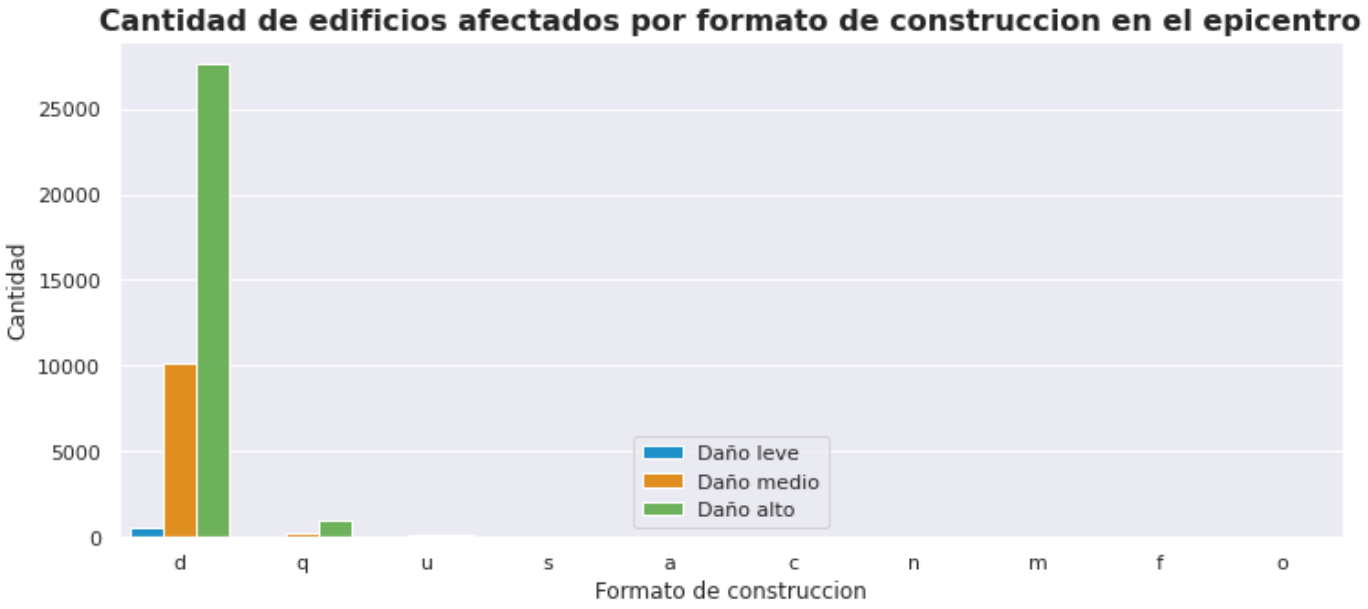

Figura 28: Cantidad de edificios dañados normalizados por nivel y formato de construcción.

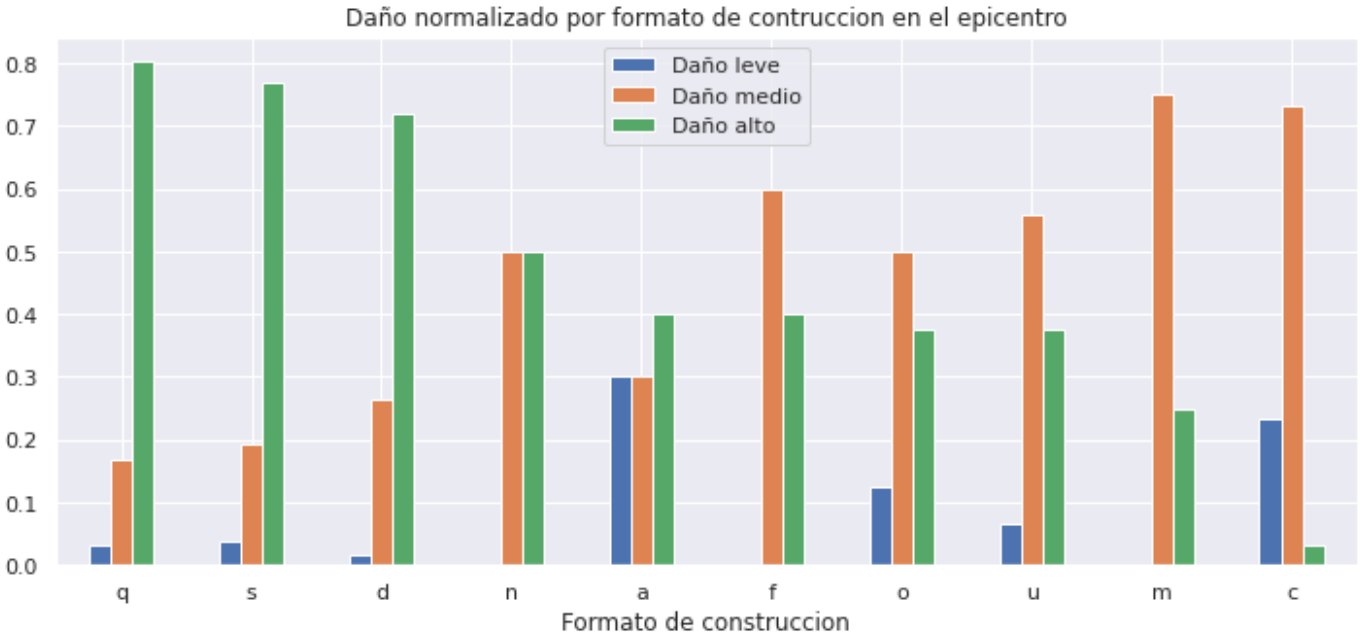

Figura 29: Cantidad de edificios dañados normalizados por nivel y formato de construcción en el epicentro.

De todas formas (tanto en el epicentro como fuera de él) se demostró que el mejor material para reducir el nivel de daño resultó ser el tipo “c” y los peores los tipos “d” y “q”.

3.3.2.7 Impacto material utilizado en PB y otros.

En este caso en particular no ameritaba realizar un análisis en profundidad del impacto en relación al epicentro, por lo cual se tomaron los valores nominales (Figura 30) y normalizados (Figura 31) para el análisis de los materiales utilizados en la planta baja de

cada edificio y por otro lado los materiales del resto de los pisos de cada edificio (Figura 32, 33 y 34).

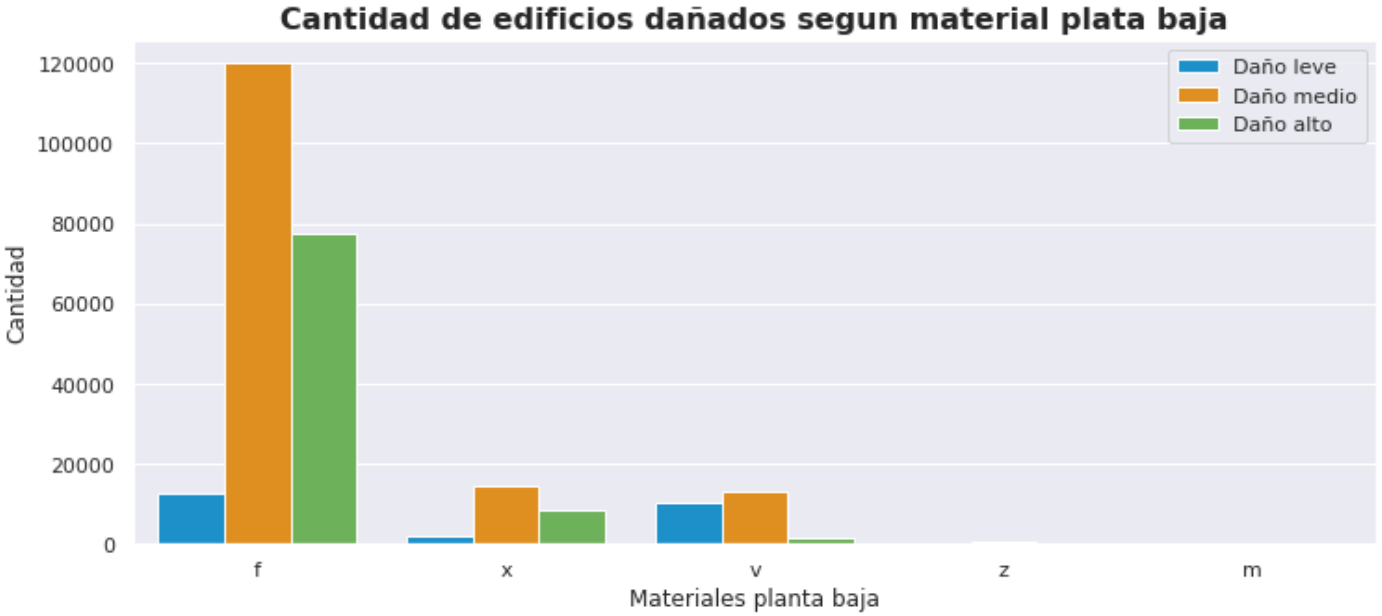

Figura 30: Cantidad de edificios dañados por nivel y material de PB.

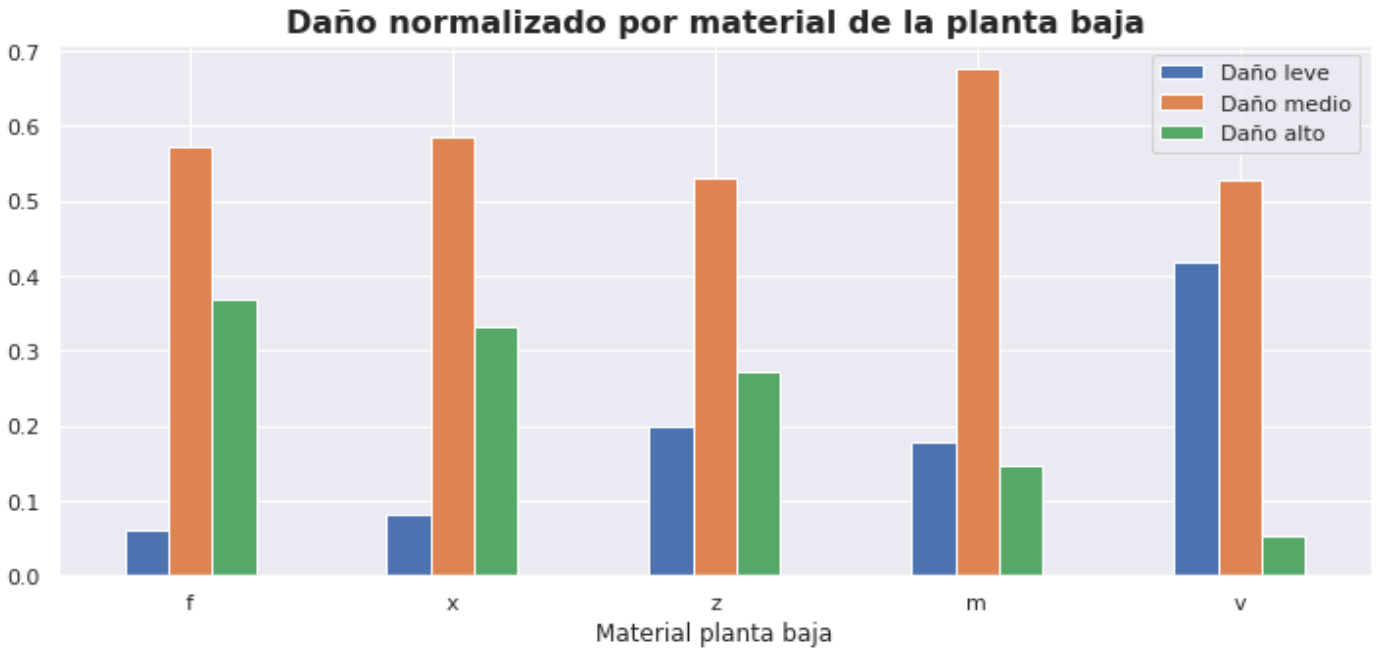

Figura 31: Edificios dañados normalizado por nivel y material de PB.

Materiales en el resto de los pisos:

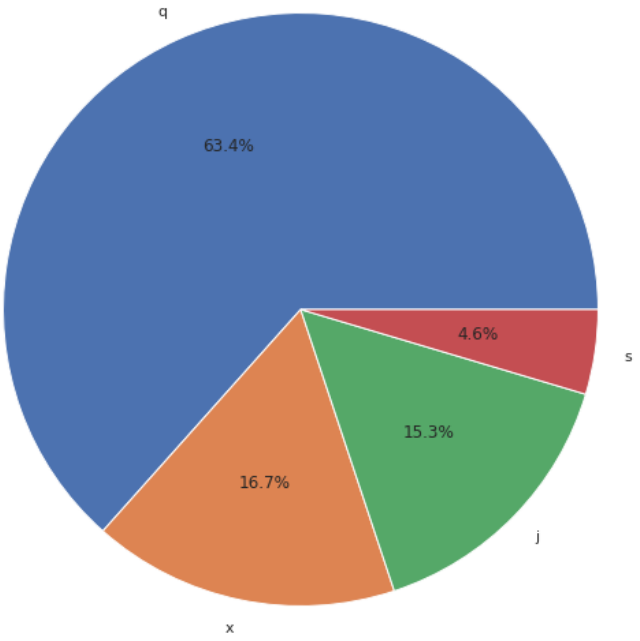

Figura 32: Proporcionalidad de daños sufridos por material de otros pisos, no PB.

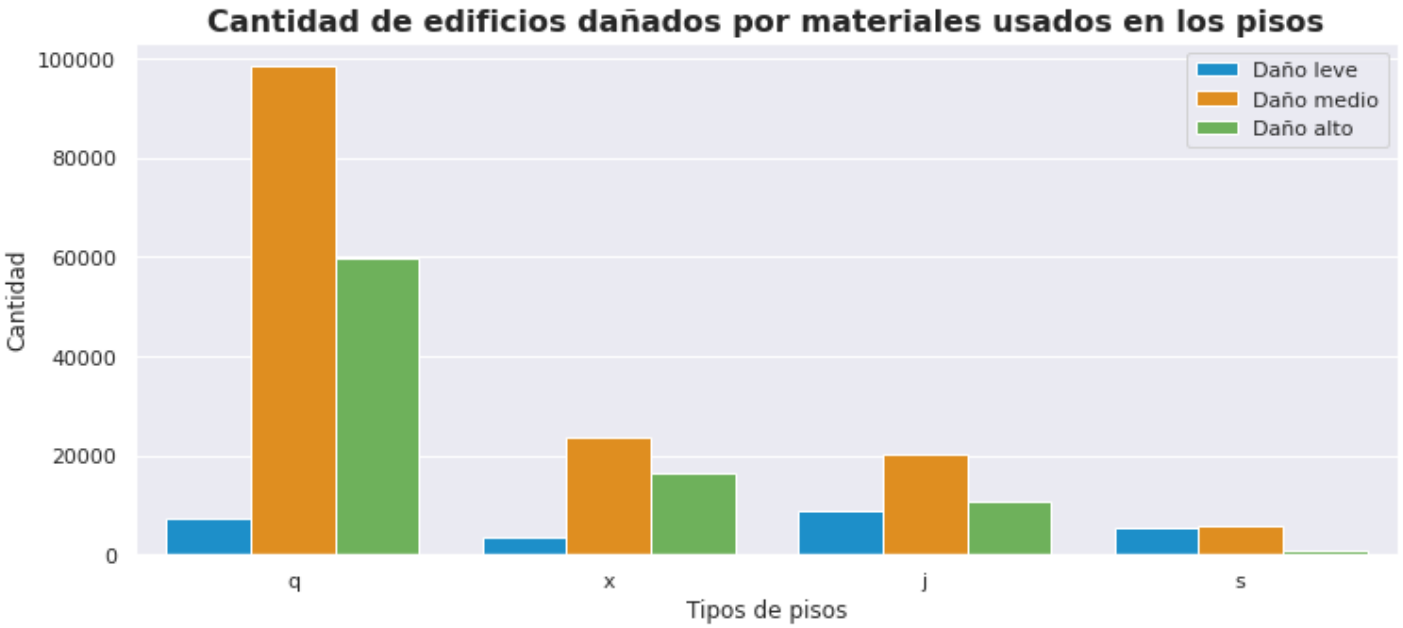

Figura 33: Cantidad de edificios dañados por nivel y material de otros pisos, no PB.

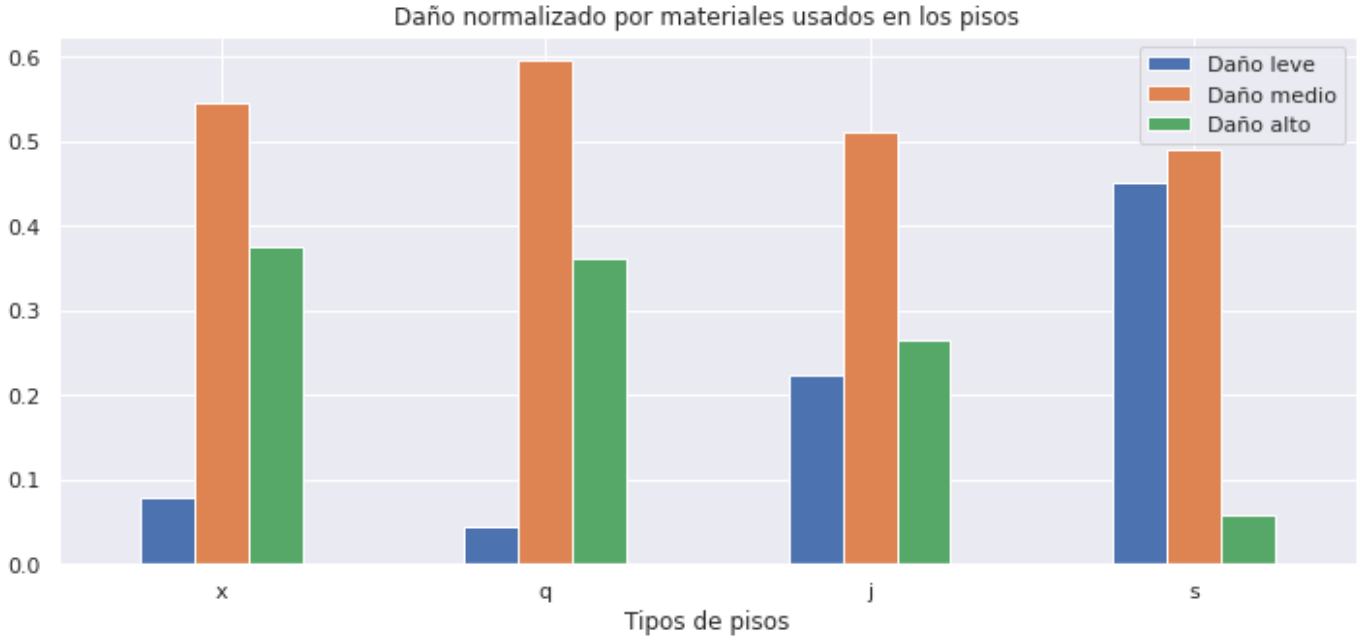

Figura 34: Cantidad de edificios dañados normalizados por nivel y material de otros pisos (no PB).

Luego del análisis se determinó que el mejor material para la construcción de planta baja es el “m”y el mejor material para la construcción del resto de los pisos es del tipo “s”.

3.3.3 Conclusión.

Del análisis realizado se detecta que existe una combinación de materiales que disminuirían el impacto del daño de una manera considerable, llevando prácticamente el daño de nivel alto a nivel bajo.
Esta combinación de materiales son:

- Estructura: Concreto Reforzado Diseñado
- Cimientos: Clase “i”
- Techos: Tipo “x”
- Formato de construcción: Tipo “c”
- Material PB: “m”
- Material pisos restantes: Tipo “s”.
- Distribución de pisos y altura: indistinto.

3.4 Impacto social (familias).

El objetivo de este punto es determinar cuántas familias sufrieron daños, en qué nivel y cómo puede evitarse o reducirse en eventos futuros.

3.4.1 Hipótesis - Preguntas.

¿Cuántas familias fueron afectadas y que nivel daño sufrieron?

3.4.2 Desarrollo.

Se analiza la cantidad de familias que sufrieron algún daño a causa de este terremoto, lo cual se visualiza en la Figura 35.

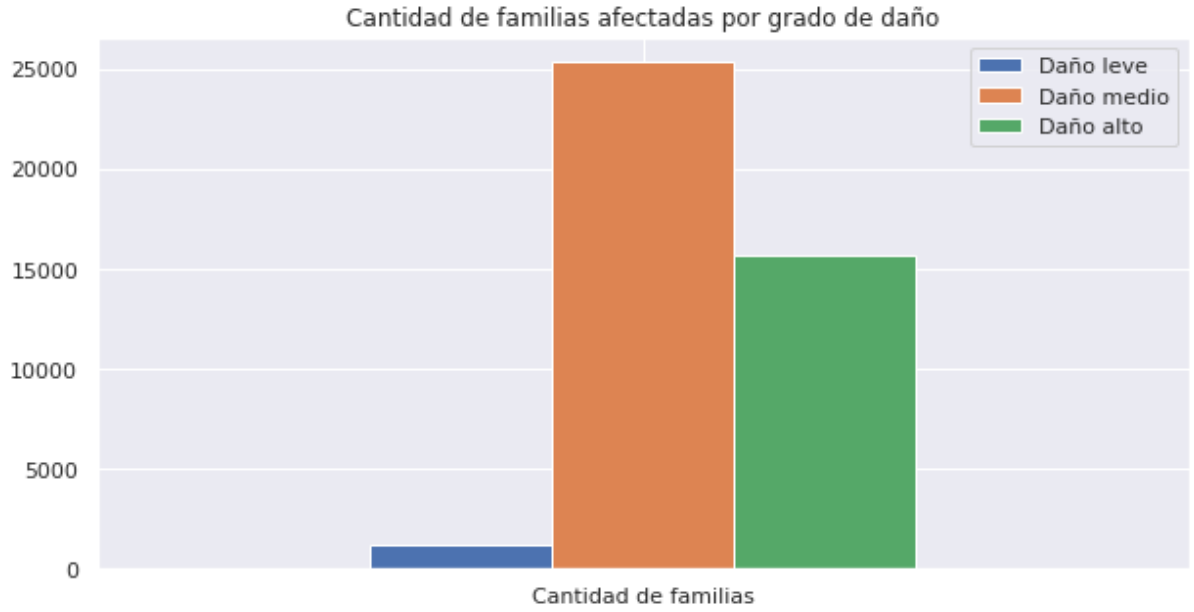

Figura 35: Cantidad de familias afectadas por grado de daño.

Se encuentran un total de 42117 familias, las cuales se dividen en los siguientes niveles de daño:

- Bajo: 1183
- Medio: 25305
- Alto: 15629

3.4.3 Conclusión.

Si tomamos la información obtenida en el desarrollo de este punto y aplicamos la lógica de la mejor construcción posible obtenida en el punto 3.3 se puede armar un nuevo gráfico (Figura 36) que demuestra cómo sería el impacto si todas las familias que poseían una de las peores estructuras, usarán la estructura elegida:

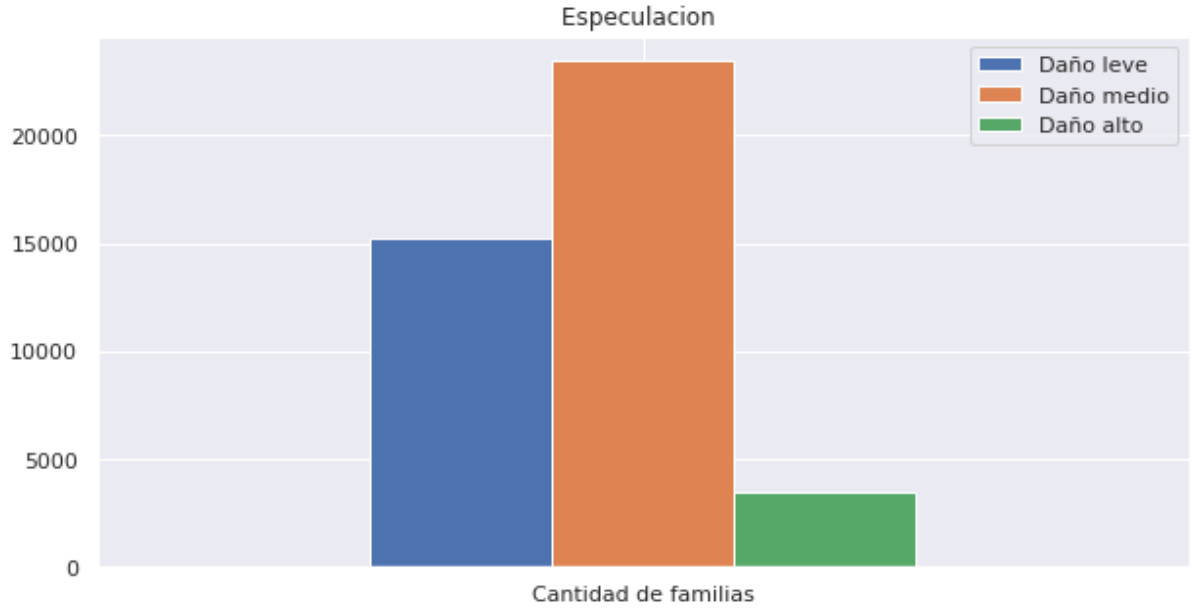

Figura 36: Proyección del daño con los mejores materiales

En valores aproximados, nos daría la siguiente distribución:

- Bajo: 15207
- Medio: 23422
- Alto: 3488

Estos valores demuestran una mejora sustancial de cara a un evento futuro de similares magnitudes.

3.5 Otros datos.

El análisis del resto de las columnas disponibles fue realizado pero no nos otorgaron patrones que puedan representar información relevante asociada a lo ya presentado.
Sin ir más lejos, algunos datos mostraban una correlación con el daño recibido pero al analizarlo en profundidad se entiende o que eran coincidencias o que se daban por la existencia de mayoría de edificios con dicha característica.
Estos campos son los siguientes:

- Area_percentage
- Land_surface_condition
- Position
- Plan_configuration
- Legal_ownership_status

4. Conclusiones finales - Insights.

Si bien nos habría sido de gran utilidad contar con algunos datos adicionales (cantidad total de edificios en el área, la relación entre los distintos niveles geográficos, entre otros) para crear comparaciones absolutas del impacto total del terremoto, mapas de daño, etc. Pudimos determinar, con la información disponible, algunos puntos importantes sobre cómo afectó el terremoto, que actividades requieren mayor recuperación cuál fue el epicentro (y por lo tanto la zona geográfica más afectada), cuántas familias se vieron afectadas y mediante la proyección de los mejores materiales y estructuras cómo se podrían reducir los daños a futuro.

A modo de resultado explícito sabemos que:

- El epicentro se produjo en las áreas 17 y 18 pertenecientes al geonivel 1.
- Los edificios de actividades de agricultura fueron los más afectados.
- La mejor estructura para los edificios se dá de la siguiente manera:
    - Estructura: Concreto Reforzado Diseñado
    - Cimientos: Clase “i”
    - Techos: Tipo “x”
    - Formato de construcción: Tipo “c”
    - Material PB: “m”
    - Material pisos restantes: Tipo “s”.
- 42117 Familias se vieron afectadas:
    - Bajo: 1183
    - Medio: 25305
    - Alto: 15629
- Proyectando los valores con la estructura óptima definida, se obtiene el siguiente resultado
    - Bajo: 15207
    - Medio: 23422
    - Alto: 3488