Detección de Anomalías en Segmento Terreno Satelital Aplicando Modelo de Mezcla Gaussiana y Rolling Means al Subsistema de Potencia.

Pablo Soligo, Germán Merkel, and Jorge Ierache

Universidad Nacional de La Matanza,
Florencio Varela 1903 (B1754JEC) San Justo, Buenos Aires, Argentina
{psoligo,jierache}@unlam.edu.ar
{gmerkel}@alumno.unlam.edu.ar
http://unlam.edu.ar

Abstract. En este trabajo exploramos la posibilidad de encontrar anomalías automáticamente en telemetría satelital real. Comparamos dos técnicas de aprendizaje automático diferentes como alternativa al control de límites clásico. Intentamos evitar, en la medida de lo posible, la intervención de un experto, detectando anomalías que no se pueden encontrar con los métodos clásicos o que se desconocen de antemano. La mezcla gaussiana y Rolling Means se aplican en la telemetría del subsistema de potencia de un satélite órbita baja. Algunos valores de telemetría se modificaron artificialmente para generar un apagado en un panel solar para intentar lograr una detección temprana por contexto o por comparación. Finalmente, se presentan los resultados y la conclusión.

Keywords: Satellites, Ground Segment, Platform, Telemetry, Machine Learning, Data Mining, Anomaly Detection

1 Introducción

El Grupo de Investigación y Desarrollo de Software Aeroespacial (GIDSA) ?? tiene como objetivo proponer y probar prototipos de soluciones de software para el área aeroespacial de nueva generación. El trabajo desarrollado incluye prototipos que utilizan interpretes de propósito general para decodificar telemetría y scripts de comandos, adopción de estándares bien probados en la industria del software, almacenamiento masivo de telemetría y detección de fallas [1], [2] y [3]. El prototipo funcional del segmento terreno se encuentra público en internet y funciona con datos de satélite reales, principalmente obtenidos de la red SatNOGS [4] y [5].
La detección temprana de anomalías en sistemas complejos como satélites artificiales son de vital importancia teniendo en cuenta el costo de las misiones y la dificultad de reparar daños. El control de límites superior e inferior para muchas variables de telemetría suelen ser la técnica más común para detectar comportamientos anómalos[6]. Como se indicó en un artículo anterior [7], la

salud del satélite se controla con la ayuda constante de un experto, utilizando poca potencia computacional. Mientras tanto, en la industria del software, el aprendizaje automático se utiliza actualmente para diferentes tipos de detección de anomalías, como fraudes con tarjetas de crédito y detección de intrusiones entre otros [8] y [9] . El objetivo de utilizar el aprendizaje automático es lograr una detección temprana de fallas evitando, en la medida de lo posible, la evaluación constante por parte de expertos así como detectar tipos de anomalías desconocidas previamente. El aprendizaje automático ofrece una interesante variedad de posibilidades de predicción y detección de anomalías. Hay dos tipos de algoritmos de aprendizaje automático: aprendizaje automático supervisado y no supervisado. El primero depende de los datos de entrada etiquetados, es decir, el conjunto de datos de entrada debe haber definido si un dato se considera una anomalía o no. El aprendizaje no supervisado no depende de los datos de entrada etiquetados, sino que aprende la representación interna del conjunto de datos y genera patrones [10]. Una anomalía es cualquier dato que se desvía de lo esperado o normal. En la literatura estadística, también se les conoce como valores atípicos o outliers. Cada dato que es procesado por el prototipo será clasificado usando etiquetas binarias: un dato es una anomalía o no [8]. Para detectar anomalías, los algoritmos de aprendizaje automático de detección de fallas crean un modelo del patrón nominal en el conjunto de datos, luego calculan una puntuación para cada valor como medida de cuan atípico es. Dependiendo del algoritmo, esta puntuación atípica toma en en cuenta la correlación con diferentes características o no [8]. En este trabajo y en el caso de series de datos de tiempo, buscamos una secuencia de valores atípicos que determina una anomalía en lugar de un dato particular, buscamos un comportamiento anormal del sistema en lugar de un valor incorrecto.
El UNLaM Ground Segment (UGS) posee el control de límites clásico desde su primera versión. En versiones posteriores se implementaron módulos prototipo que modifican los limites dinámicamente[7]. En este trabajo, en lugar de trabajar con límites, buscamos obtener una medida de éxito en la detección de un comportamiento anómalo del subsistema de potencia, comportamiento que no puede ser detectado por el control de limites clásico. El trabajo actual se exploran dos métodos de aprendizaje automático diferentes mezcla gaussiana y rolling means. Estos dos algoritmos son investigados y comparados entre sí para estudiar la viabilidad de aplicarlos en la detección de patrones y comportamientos en un prototipo de control de salud en tiempo real.

2 Materiales y métodos

Para estos experimentos usamos nuestro propio conjunto de datos con datos de telemetría real [11]. La fuente de la telemetría es un satélite científico de órbita baja. La telemetría comienza en 2015-05-27 08:51:06 +00:00 y termina en 2015-06-05 23:34:06 +00:00. Para estos experimentos usamos solo los dos primeros días, desde 2015-05-27 08:51:22 +00:00 hasta 2015-05-29 08:50:59 +00:00. El conjunto de datos de entrenamiento finalmente tiene 17277 tuplas, el conjunto

de datos de prueba tiene 4320 tuplas. La tabla 1 muestra el significado de cada campo según descripción disponible en la documentación del fabricante. Todos los valores, excepto *vBatAverage* y *BatteryDischarging* están en bruto(raw), sin embargo, los datos siempre se normalizan antes de ser procesados. Desafortunadamente, el conjunto de datos no cuenta con fallas documentadas por un experto. Para crear una anomalía artificial, cortamos parcialmente la generación energía del panel solar 24 poniendo en 0 la corriente (128 en bruto) en el conjunto de datos de prueba. El corte es progresivo y cubre 1079 tuplas. Esto es similar a dejar el panel eclipsado (según posición orbital), independientemente del contexto real. Tenga en cuenta que el límite clásico el control no puede manejar este comportamiento, debido al hecho de que las corrientes cercanas a 0 son perfectamente válidas en períodos de eclipse.

Feature	Meaning
vBatAverage	Average of Battery voltage used by supervisions
BatteryDischarging	Flag True/False if battery is discharging
ISenseRS1	IsenseRS1 current (battery current)
ISenseRS2	IsenseRS2 current (battery current)
V_MODULE_N_SA	Current in solar panel #N con 0<N<25

Table 1: DataSet Features

Se utilizan dos algoritmos de aprendizaje automático diferentes para detectar anomalías en Telemetría satelital: Mezcla gaussiana y Rolling Means. El primero se aplica la telemetría del subsistema de potencia en su conjunto, utilizando la correlación entre variables, mientras que la última se aplica a cada variable de telemetría de manera aislada. Ambos modelos siguen enfoques estadísticos clásicos: ambos utilizan medidas estadísticas como media, desviación estándar y probabilidad.

2.1 Gaussian Mixture Model

Todas las telemetrías del subsistema de potencia están altamente correlacionadas como se muestra en la figura 1, por razones de tamaño, mostramos la correlación de solo 4 de las 24 características actuales del panel.

Usando la biblioteca sklearn [12], creamos una Modelo de Mezcla Gaussiana, del Inglés Gaussian Mixture Model (GMM). GMM puede utilizarse para agrupar datos sin etiquetar, GMM puede ayudar a detectar un comportamiento lejano o poco probable que el comportamiento nominal. Cualquier punto muy alejado de las funciones gaussianas podrían considerarse una anomalía. El conjunto de datos se dividió en dos subconjuntos de datos: conjunto de datos de entrenamiento y conjunto de datos de prueba . La prueba se realiza durante dos días de telemetría. El 20% final del conjunto de datos se utiliza para la prueba mientras que el otro 80% forma el conjunto de datos de entrenamiento. Un modelo es obtenido al

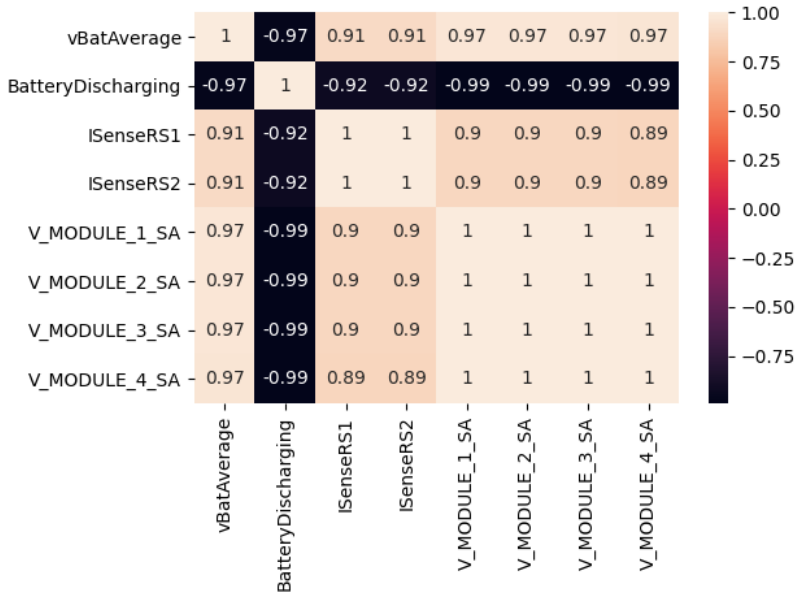

Fig. 1: Correlaciones entre valores de telemetría del subsistema de potencia

ejecutar el algoritmo sobre el conjunto de entrenamiento, donde la cantidad de componentes y el tipo de covarianza se seleccionan en un proceso iterativo que analiza information-theoretic criteria (BIC), cubriendo los 4 tipos de covarianza y la cantidad de componentes entre 1 y 20. La puntuación mínima de valores atípicos se establece como límite para la prueba.

2.2 Rolling Means

Rolling Means utiliza un enfoque estadístico simple para la detección de anomalías sobre un conjunto de datos de serie de tiempos. Dada una serie de referencias y una ventana de tamaño fijo N, el algoritmo obtiene primero la media de los N registros iniciales de la serie. Entonces la ventana se "mueve hacia adelante" en uno, recalculando la media de la ventana. Este proceso se repite hasta que la ventana final incluye el dato final. Una vez que todos las medias se ha obtenido, el algoritmo etiqueta como valores atípicos todos los puntos cuyo desvío de la media es S veces mayor que la desviación estándar que corresponde al punto.

Se aplica Rolling Means mediante un algoritmo [13] a cada subconjunto de datos, uno para cada variable de telemetría, y para cada uno genera un modelo de normalidad. Para utilizar este algoritmo, se debe establecer el tamaño de la ventana y un número fijo de desviaciones estándar. Para decidir el valor de estos parámetros, se ejecutan varias iteraciones con diferentes valores en el mismo conjunto de datos, y finalmente, conociendo la naturaleza de los datos y tomando el rol de experto, los mejores valores se utilizan.

Se eligió Rolling Means ya que es un algoritmo sensible a valores anómalos, siendo simple de implementar. Se basa en la desviación estándar, teniendo en cuenta el cambio en la serie de tiempo usando la ventana de tamaño fijo.

2.3 Otros métodos

También se probó el método de distribución normal multivariable, pero se descartó a favor de la Mezcla Gaussiana, dado que el primero necesita que sus datos sigan una distribución normal y no puede manejar varias campanas. Isolation Tree también fue analizado, pero se descartó dado que se etiquetaron incorrectamente la mayor parte de el conjunto de datos "normal" como anomalías, sin tener la posibilidad de utilizar un parámetro para cambiar su comportamiento.

3 Resultados

3.1 Modelo de Mezcla Gaussiana

Para obtener un grafico que nos brinde una aproximación visual al modelo generado utilizamos inicialmente solo dos características $V\_MODULE\_24\_SA$ y $vBatAverage$ sobre los datos de entrenamiento.La figura 2 muestra graficamente las funciones gaussianas en verde y las diferentes agrupaciones generadas.

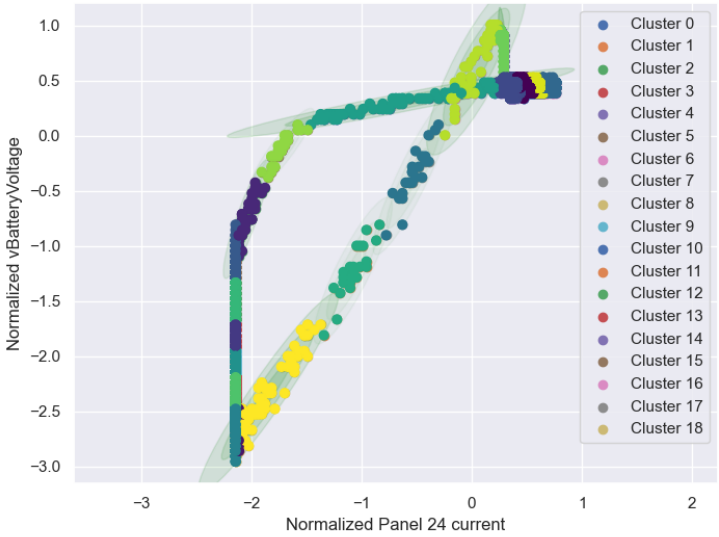

Fig. 2: Grupos o Clusters creados para 2 variables usando mezcla gaussiana

Se testean los datos de prueba con el modelo previamente generado. La figura 3 muestra como los datos de prueba sin modificación artificial ajustan al modelo.

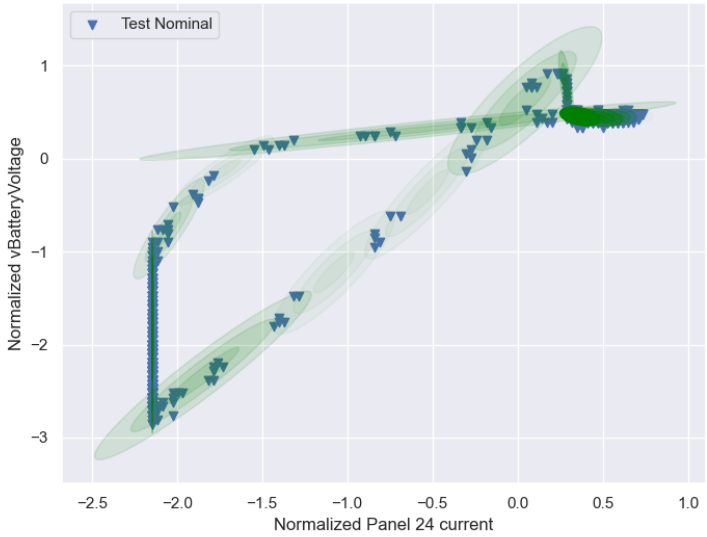

Fig. 3: Dataset de prueba sin anomalías generadas artificialmente. En verde las funciones gaussianas, todos los datos se ajustan al modelo. No hay falsos positivos

Los resultados con el conjunto de datos modificado artificialmente, con solo 2 variables ($V\_MODULE\_24\_SA$ y $vBatAverage$), simulando corriente 0 en el panel solar 24, se muestran en la figura 4. Se detectan 880 anomalías. Si bien la caída progresiva de corriente no permite separar de forma claro cual dato es anómalo y cual no, la cantidad obtenida sobre el total de datos es una buena medida del estado general del sistema.

Si utilizamos las 28 características disponibles en el dataset 1 obtenemos 11 falsos positivos con el conjunto de datos sin modificar, es el 0,25% del conjunto de datos de prueba, si, también usando todas las características 1 pero con el dataset modificado artificialmente obtenemos 925 anomalías. Los resultados del experimento con 2 y 28 características, para datos originales o modificados artificialmente son mostrados en la tabla 2.

3.2 Rolling Means

El conjunto de datos se divide en 7 subconjuntos de datos, uno para cada variable de telemetría. Para cada uno de estos subconjuntos de datos, el algoritmo Rolling

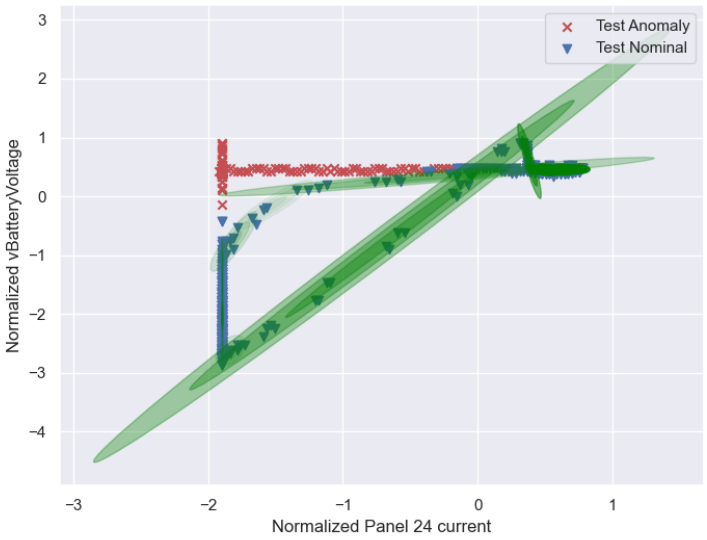

Fig. 4: Dataset de prueba con anomalías

# Variables	Dataset Normal	Dataset c/anomalias
2	0	880
28	11	925

Table 2: Anomalías detectadas por GMM

Means etiqueta los datos de cada variable como anomalías o no, según el "modelo de normalidad".

Ejecutando el algoritmo, con un tamaño de ventana de 1000 (una ventana que es la mitad del número de anomalías insertadas), utilizando 1 y 2 desviaciones estándar obtiene los próximos resultados. Cada subconjunto de datos se traza con líneas azules que representan datos considerados nominales y lineas rojas que representan los puntos de anomalías que detectó el algoritmo. La tabla 3 muestra, para una y dos desviaciones estándar la cantidad de anomalías detectadas en el dataset original y el modificado artificialmente.

4 Conclusiones

En el caso de Rolling Means, el modelo no es sensible a la correlación y define si un dato es anómalo basándose en la tendencia del conjunto de datos de series de tiempo. Un pico aislado en el gráfico se etiquetará como una anomalía, pero

#desviaciones estándar	Dataset Normal	Dataset c/anomalías
1	260	449
2	71	6

Table 3: Anomalías detectadas por Rolling Means para una y dos desviaciones estándar sobre dataset original y modificado

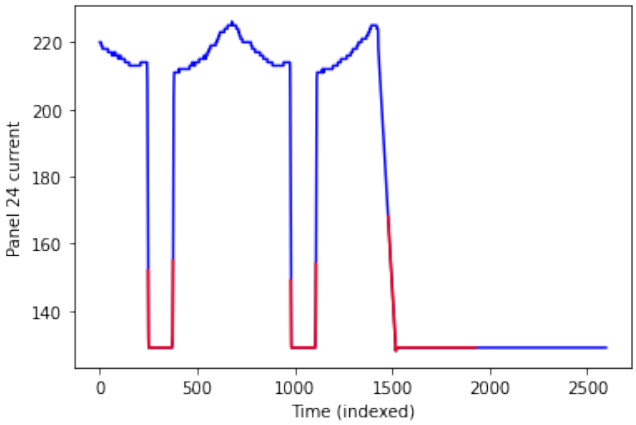

Fig. 5: Rolling Means aplicado al Panel 24 usando una desviación estándar

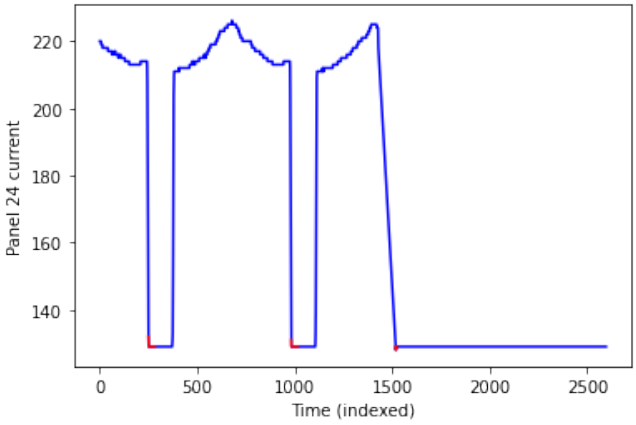

Fig. 6: Rolling Means aplicado al Panel 24 usando dos desviaciones estándar

puede ser el resultado de una acción contextual esperada. Aunque Rolling Means es un algoritmo simple, con pocas necesidades computacionales, al no tener en cuenta el contexto y las correlaciones no pueden manejar anomalías específicas y

dependientes del contexto. Usando una desviación estándar parece detectar las anomalías introducidas, pero también etiqueta erróneamente los datos válidos.

Usando dos desviaciones estándar, contrariamente a lo esperado, se comporta de la misma manera, etiquetando incorrectamente aún más datos. Rolling Means es un método válido para detectar valores atípicos producidos por ruido, pero no puede considerarse un algoritmo válido para la detección de anomalías. También necesita la intervención de un experto para establecer los parámetros iniciales. Por otro lado, el método de Mezcla Gaussiana muestra resultados prometedores. Se detectaron anomalías, sin etiquetar incorrectamente una gran cantidad de registros (signo de que el modelo no se ha sobreentrenado). Estas anomalías introducidas no pueden ser detectadas por los sistemas de control de límites dado que los valores probados son normales en un contexto determinado. La covarianza y la cantidad de clusters se obtuvieron automáticamente, sin un intervención experta.

5 Trabajo Futuro

Los resultados dan una vista informativa de los diferentes algoritmos, pero no pueden ser evaluado objetivamente ya que no hay datos etiquetados disponibles para compararlos. Si se pudieran obtener datos preetiquetados, se utilizarían métricas estadísticas para evaluar los resultados y ajustar los parámetros de los modelos para minimizar el número de falsos positivos producidos por el prototipo.

Entre los requerimientos típicos de estos sistemas se encuentra la detección de anomalías. Otros algoritmos como DBScan, y técnicas de aprendizaje profundo deben ser exploradas a futuro, la correlación de variables ha demostrado ser un capital valioso para detectar comportamientos anormales y pueden ofrecer una solución superadora al simple control de límites. Lamentablemente no es posible a la fecha encontrar un conjunto de datos satelital real con anomalías documentadas o fallos futuros que podrían detectarse mediante un análisis histórico de datos. Este limitación es una de las mayores barreras a vencer en trabajos futuros si se pretende alcanzar en este área el mismo nivel de implementación que en otros sectores (Fraude Bancario, Intrusiones, Diagnósticos Médicos, etc).

References

1. Pablo Soligo and Jorge Salvador Ierache. Software de segmento terreno de próxima generación. In *XXIV Congreso Argentino de Ciencias de la Computación (La Plata, 2018).*, 2018.
2. Pablo Soligo and Jorge Salvador Ierache. Segmento terreno para misiones espaciales de prÃşxima generaciÃşn. *WICC 2019.*
3. Pablo Soligo, Jorge Salvador Ierache, and German Merkel. Telemetría de altas prestaciones sobre base de datos de serie de tiempos. 2020.
4. Unlam Ground Segment: Home unlam ground segment: Home. `https://ugs.unlam.edu.ar/`. Accessed: 2021-07-30.

5. Satnogs satnogs. `https://satnogs.org/`. Accessed: 2021-07-30.
6. Takehisa Yairi, Minoru Nakatsugawa, Koichi Hori, Shinichi Nakasuka, Kazuo Machida, and Naoki Ishihama. Adaptive limit checking for spacecraft telemetry data using regression tree learning. In *2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No. 04CH37583)*, volume 6, pages 5130–5135. IEEE, 2004.
7. Pablo Soligo and Jorge Salvador Ierache. Arquitectura de segmento terreno satelital adaptada para el control de límites de telemetría dinámicos. 2019.
8. Charu Aggarwal. *An introduction to outlier analysis*. Springer New York, 1 edition, 2017.
9. Aaron Rosenbaum. Detecting credit card fraud with machine learning. 2019.
10. Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of Statistical Learning*. Springer Series in Statistics. Springer, 2 edition, 2008.
11. Low Orbit Satellite Dataset: Home low orbit satellite dataset: Home. `https://gidsa.unlam.edu.ar/data/LowOrbitSatellite.csv`. Accessed: 2021-07-30.
12. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
13. Algoritmo Rolling Means rollingmeans. `https://gidsa.unlam.edu.ar/data/rolling.py`. Accessed: 2021-07-30.