**BART Rider Commuter Patterns**

Kenneth Ceglia

Colorado State University Global

Course Code: MIS581

Instructor Dr. Ford

Due Date 4/18/2021

**BART Rider Commuter Patterns**

People form patterns. Patterns are seen in almost every aspect of daily life. Many people wake at the same hour, leave for work and follow the same path each day to and from work. This is not entirely uncommon with other communities. Bees travel the same paths to and from their hive, ants seek out and establish new, repeatable paths. In larger cities with extensive public transportation networks, such as BART or “Bay Area Rapid Transit” many people commute on the same route lines each and every day, with surprisingly regularity.

The rapid transit system in the Bay Area of California has approximately 50 stations (expanding regularly) and 8 major lines. Close to half a million people ride BART every day through a network of over 2500 different combinations of stations. Predicting where riders are heading each day opens the door to improved rider services, improved health conditions and new revenue streams through direct to rider marketing and advertisements.

The BART rail system resembles a living network that grows and shrinks and changes each hour. In the morning, commuters follow certain patterns and routes. Holidays and special events may temporarily change the commuting patterns, but offer new opportunities for improved rider services and marketing opportunities for local business to reach a larger number of potential patrons at very predictable times. Knowing when and where riders are traveling provides enhanced opportunities for local business to advertise services and offerings directly to consumers with predictable regularity and time exposure.

While many public transportation services offer ridership numbers, it is less common to see a rider distribution breakdown of where and when and what paths riders takes each day (and hour). The focus of this research is to discover these ridership route patterns and if those patterns are repeating with regularity and predictability.

**Objectives**

The focus or objectives of this research is determining if route patterns are discoverable, repeatable and predictable and if these patterns provide opportunities for rider and local business services. There are several different types of patterns in this study.

When looking at ridership in a given rail or subway system, there are several levels or granularities to consider. On one end is total ridership. How many riders are using the rail system? Knowing if total ridership is changing is particularly important for the BART system itself. For example, knowing if total ridership is increasing or decreasing can direct focus towards improved rider services, such as more trains or better schedules. Total ridership over time is also useful in understanding seasonal fluctuations, such as more riders in the winter months over summer months. Transportation services may use this information to offer increased services during up cycles and possibly decrease services or possibly add incentives to encourage more riders in down seasons.

The next level considers ridership at individual stations. For example, knowing the total number of riders that queue at a single station each day provides a more targeted objective. Further, knowing how many riders queue at a station each hour provides a finer detail for rider services. Further yet, knowing the number of riders departing a station and the distribution of riders to their destinations provides extremely high-fidelity detail of ridership patterns. Services may be offered that are targeted for the distribution (and number) of riders heading for a known destination. This detail provides local business the opportunity to offer goods and services to a predictable number of patrons arriving by their local businesses, even at a given hour of the day. Services or marketing may be offered at the departing station and continued in transit (on train) knowing confidently the number of riders and duration to their arrival station.

The objective of this research is to discover the ridership patterns that focus on high level and detailed route patterns over time and the predictability of those patterns that can be used by BART services, local businesses and improved health and sanitation services.

# **Overview of Study**

BART is an advanced public transportation network that links communities in the San Francisco Peninsula with cities in the Eat and South Bay as far south as Millbrae and as north as Richmond. BART has been in operation for over 45 years carrying citizens to and from work, universities and colleges, special and sporting events. BART runs on all electric cars with the majority of power (95%) coming from renewable energy sources making BART one of the most environmentally successful public transportation systems in the world (*Sustainability*, n.d.).

Like any service, BART faces competition and demands for improved services. Over the last few years (prior to Covid-19) ridership has declined due to crowded trains, security and sanitation concerns. Riders expect more services, better schedules and improvements to stay on board. One of the ways to improve BART is to better understand who, when and how it is being used.

The goal of this study is to analyze ridership route patterns to provide analytics that can be used to improved the lives and services of BART riders and local businesses (Wang et al., 2019). By understanding and predicting the ridership profiles, BART and local businesses can offer improved services to attract new riders and retain existing ones (Andrews et al., 2016).

Commuter route patterns are not normally provided by subway or rail services. BART also does not provide explicit route pattern data. This study aggregates data from multiple BART repositories and route schedules to create these route patterns. The route patterns are synthesized in a format with a maximum resolution of 1 hour, which provides a high granularity of entry/exit numbers per station, per hour.

With 50 stations, there is a possibility of 50x49 or 2450 different routes each hour. It is not practical to analyze each route, however a predictive model may be produced that can be applied to different routes to predict the number of riders entering/existing at each hour, provided the data is in a format that is easily queried, such as a relational database.

This study will first look at overall ridership numbers and find the most used, or highest traffic stations, since those would be assumed most interesting. It is likely these are stations in the surrounding cities where most commuters tend board from. Once the origin stations are identified, several routes are selected to analyze a profile of rider routes. As mentioned earlier, it is not practical or even feasible to look at all routes, so the most traveled or highest used routes will be identified and used as the basis for the study.

Once the routes are identified, the ridership distributions are analyzed to determine how repeatable these routes are. How and if they change over time and to what level of granularity can ridership distributions are predictable. This involves a bit of inductive research as these answers are the results of analysis. Deductively, this study looks to prove (or disprove) the assumption statistically, that routes are repeatable and ridership patterns are statistically significant. To which stations, hours and granularity, such as days or weeks or months, is inductive to the research.

One side element of research is the impact of Covid-19 on ridership data. It is well known that the number of subways riders has drastically reduced. BART reduced schedules and shut down multiple train services once the pandemic hit the Bay Area. Fortunately, there is an abundance of historical data that will be used in this study going back to 2013 up to 2021. Mixing pre and post Covid-19 data may not be possible given the dramatic change in ridership and train schedules.

If the ridership routes distributions are repeatable and statistically significant, a final model, such as recurrent neural network or other predictive algorithm, can be considered to predict future ridership. Mobile Apps or web sites may potentially incorporate the rider prediction numbers to give riders queue or crowd size estimates. Local business can use such sites to estimate the number of patrons “passing by.” Lyft and Uber drivers may also consider adopting such tools to estimate the number of riders existing a station and target their services towards the most likely patrons.

BART may use estimates to offer marketing or advertisement incentives to business. These advertisement services can be offered at stations or onboard with an estimated target population and advertisement exposure duration.

**Terms**:

**BART** – Bay Area Rapid Transit. The heavy rail commuter rail system serving the Bay Area.

**Route** – this is the path a rider takes. A rider starts at one station and arrives at another. A route is a transactional “start to stop” event.

**Origin** – station that riders depart from

**Destination** – statin that riders arrive at

**RRD** – Ridership route data. This is the number of riders starting at one station and arriving at multiple stations. This is the distribution of riders and their routes from a departure station.

# **Research Hypothesis**

*BART Riders follow the same and predictable network route patterns for each day and hour that can be used to improve services, direct local businesses advertisements and increased rider safety*.

**RQ1**: Is the number of riders in queue at a station related to the day of week and time of day?

The NULL hypothesis is that the day and hour have no effect on the queue of riders for a station. The alternative is that the day and time do have an effect on the number of riders in queue for a given station.

**RQ2**: Do the same proportion of riders go to the same stations each day for morning commutes for a given day of the week?

: =

They NULL hypothesis is for any day of the week, such as Monday or Friday, that the proportion of riders that go to the same station each morning is the same.

### 

**RQ3**: Are the number of riders arriving and existing from a station related to the day of week and hour of day?

**RQ4**: Has the total number of BART riders decreased in recent years?

The NULL hypothesis indicates the number of riders has remained the same, or a regression fit indicates the slope is constant. The alternative is the regression slope is negative, indicating ridership has decreased. Note: This is assuming a linear fit, it may end up an exponential or quadratic fit, where ridership may have increased, then decreased recently. In this case, either a piece-wise linear may make sense, or simply test a negative quadratic coefficient.

# **Research Design**

The following discusses the research methodology and how the study is conducted start to finish, the logic and sampling methods used. Included in the research design is the methods used, which specify how the data is collected and processes followed by statistical tests used to validate or disprove hypothesis that riders are predictable

## **Methodology**

This primary analysis is quantitative. The Ridership data, schedules and trip times required for the study are exclusively numeric. The analysis for this study is quantitative, using classical statistics methods to test hypothesis, and provide predictive outcomes. Machine learning approaches may be used as well for predictive responses. Below looks at the methodology used throughout the study to validate (or reject) the hypothesis that rider routes are determined by the day of week and hour of day.

**Process Flow**

The methodology flow starts with data extraction, data reduction, exploration. These three steps are mostly linear with some iterations between reduction and exploration. Following are hypothesis testing. Lastly, predictions methods are explored using machine learning models, such as neural nets or classical statistical methods, such as seasonal ARIMA models.

**Logic**

The research type used is a basic research framework that uses a deductive approach to test the theory that riders following predicable route patterns based on day and hour. There is a small inductive element that seeks to discover the highest traffic routes. Not all routes are likely traveled extensively with high volumes of riders. There are likely some routes with very small riders and possibly random with no correlation. These are also important to observe. Given the number of stations and permutations, it is likely some inductive logic will be used to find the highest and lowest and “moderate” travelled routes out of the several thousand possible.

**Sampling**

Secondary sampling is used. The samples used are sourced from BART Open Data Site where the data has already been collected. This has some implications when testing – specifically for route proportion comparisons using a c chi square test for homogeneity. These tests assume data has come from the same population. For this study, that is the assumption, but unproven or verified. It is assumed the same population is riding BART each day with occasional exceptions.

**Data Aggregation**

Data is first extracted from BART’s Open Data site and route data is aggregated using a combination of ridership entry/exit data and schedule data over a period of 2013 to 2021. The details of extracting and aggregating the data is contained in the “methods” section below.

Once aggregated and the required RRD data is synthesized, and the resulting data is placed in a Postgres SQL database where it is easily extracted and explored. The quantity of data is very large. The number of BART stations is between 40 and 50 through the 2013-2021 period. The number of route permutations with this number is exceeding large, over 100 million records.

**Station and Line Selection**

To reasonably sample and explore the data, the data is reduced to a single route line and only weekdays are considered. The **Pittsburgh to San Francisco** is planned for selection to limit the number of routes to explore. Once reduced to a single line is selected, the highest traffic route, such as **Pittsburg to Embarcadero** is explored to begin testing data with. This reduces the data to around ~50 million records.

**Environment**

Python is the language used in the study, due to its flexibility and tight integration with PostgreSQL. The data set is very large, over 50 million samples or 3.5 GB of data. To explore and manipulate this data, a fast database and language capable of handling this size of data efficiently is needed. Since rider data is in CSV and schedule data in HTTP queries, a language capable of handling both types easily were needed. Data tables are fully indexed on primary keys and combinations, greatly speeding up data access (a typical desktop computer is adequate).

**Repo**

All code is stored at: [KPZNet/MIS581Capstone (github.com)](https://github.com/KPZNet/MIS581Capstone)

**Visualizations**

Python ***Matplotlib*** were chosen for visualizations due to ease of use, popularity and swift integration with PostgreSQL.

**Statistical Tests**

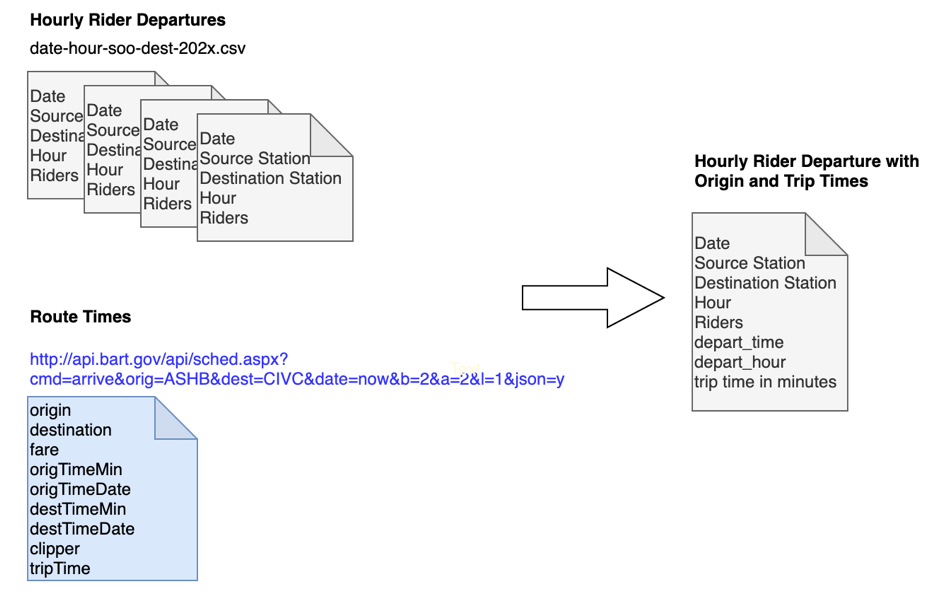
Python ***SciPy*** and ***Statsmodels*** packages were selected due to ease of use, popularity and integration with PostgreSQL.

**Methods**

Quantitative analysis is the primary method used. The study focuses on ridership data and rail schedules, which in both cases are exclusively numeric. The data samples are provided by BART open data sources and contained in several individual CSV formatted files. Quantitative data is provided in multiple resolutions over several tables. Schedule data is provided through HTTP interface from BART Open Data Initiative.

To analyze the RRD distributions, the ***Schedule*** and ***Ridership Data*** is aggregated or joined to create a new quantitative data repository that contains trip times for each route in the network. The following describes the process used to aggregate route distributions.

**Ridership Data**



BART provides Entry/Exit data for each year starting in 2013 through 2021. Each year is contained in a single CSV or comma separated value table that includes the following detail:

Date | Hour | Entry Station | Exit Station | Riders

Per Hour. These tables are very large ~ 500MB per year with 8 years’ worth

This format contains the date and hour and number of riders that “exited” that station for all routes taken by 1 or more riders per hour.

An alternative view is that this is the number of riders “arriving” at the Exit station from the Entry station. Important to note that this number is total number of riders arriving and existing that station. It does not include when the riders boarded.

**Schedule Data**

The schedule table is acquired from BART through HTTP interface based on requested trip, date and time of day, and contains the following detail.

Date | Time | Origin | Destination | Minutes

NOTE: This includes the complete trip duration – from the Origin to the Destination in minutes. This includes the estimated time between transfers. Since transfers are dependent on multiple schedules, BART estimates the total trip time based on number of trains and estimated departure intervals for each train at the requested time of the day.

Note: *Since trip times are dynamic per time of day, and trip times for historical ridership data is not provided, trip times were estimated using current schedules “averaged” over each day and hour. This gives us one trip time for each route. The variance of trip time was small and generally fit into the 1-hour resolution of the ridership data. The ridership data is always bucketized to 1 hour, which makes absolute departure times only an estimate, but consistently within the arrival time 1 bucket, or 1 hour behind.*

**Ridership Route Data**

In order to create a RRD profile, the ridership data and schedule data are aggregated to add “trip time” and “originating” time to the ridership data. This new aggregate format gives the following detail:

Date | Hour | Entry Station | Exit Station | Depart Date | Depart Hour | Minutes | Riders

This aggregate data now provides when the riders departed, and duration of trip.

**Statistical Significance**

To test RQ1 and RQ3, the ridership data is analyzed using classical statistical methods using Time Series analysis. To estimate total ridership arrivals or queue lengths, time series analysis is used with auto correlation and/or partial auto correlation to predict the queue and station congestions and total riders (RQ1, RQ3).

To test RQ2, rider commuter route patterns, the number of riders or distribution of riders route patterns, a c chi square test of homogeneity is used to compare rider distributions to determine if there is a consistency between proportions of riders following the same routes between different weekly, or monthly rider totals.

To test RQ4, a linear regression or exponential regression is planned to determine if there is a trend and what direction is that trend. Is it declining indicating a reduction in total riders?

Data cleaning is done as close to the tests as possible. There are several events that can change the data collected. For example, holidays, strikes, special events such as sporting events or other high-capacity events. Since these events are not errors, they are important data points, rather than scrub them out early, special events remain in the data and are “filtered” out as needed based on the analysis.

Some other events are scrubbed early and removed, such as incomplete trips or fraudulent “line jumpers” where the rider skipped or jumped the turnstile to avoid payment. These events were scrubbed early to avoid any influence.

# **Limitations**

One immediate limitation is the number of routes to consider. To facilitate reasonable exploration early, only a subset of routes is considered at first. This is due to extremely high number of routes possible. Another limitation is the data granularity. The RRD is bucketized at 1 hour. While normally, this would be considered high fidelity data, it does summarize the data into only single hours of total riders which does introduce some errors when synthesizing or joining route data with schedule data.

Covid-19 presents a complication with normal rider data. After the pandemic hit the Bay Area, BART ridership drastically decreased and schedules and availability were reduced. This could throw off the hypothesis or purpose of the study so Covid-19 months are excluded from the study. This creates a limitation, or a possible one. The study is focused on determining if riders travel in patterned routes. It is entirely possible that the same route patterns still exist during Covid-19 periods. In fact, this would only strengthen the study, showing route patterns are more resilient than expected. However, it is possible that Covid-19 data is removed or treated separately in this study than pre-Covid ridership numbers.

# **Ethical Considerations**

All data used in the study is acquired from BART directly through the Open Data Initiative. There is an assumption here that the data has been carefully scrubbed for any personal details or other privacy details embedded in rider data. Still, an abundance of caution and attention is made to ensure any details that might has gone unnoticed are not exposed in this study. Anything to the contrary will be made known to BART authorities.

There is always a concern how somebody may use this data in a negative way. This is probably always the case with any research. Somebody can use the results for wrong-doing. This study focuses (or attempts to) on the good or positive outcomes possible through better understanding of rider commute routes and station utilization. Ideally, this, if anything at all, will encourage end users to focus on these positive potentials.

***Conflict of Interest***

This study does represent a small conflict of interest in that the author uses the subway system as the primary means of personal transportation. There is a natural desire to keep the subways improving, so there could be a bias towards seeking those results.

# **Literature Review**

Studying subway or light rail usage is an ongoing effort for most major metropolitan cities. One of the best ways towards continual improvement is measuring and studying the current usages and finding new ways to improve based on how riders are using the rails. New stations are added, new cars are added or removed based on rider utilization. Like any business, subways or light rails go through up and down cycles. High crime, poor sanitation, high costs or simply better alternatives can sway commuters to other forms of transportation. While subways and light rails are forms of public transportation usually managed by government authorities, they still face the same competition as any business and must continue to innovate and find new ways to improve services to stay competitive and healthy and reduce congestion (Keler et al., 2020).

One simple way to keep services competitive is timeliness. Late trains, overcrowded stations are often seen as the biggest deterrent from using subways. Riders would rather endure impacted trains than late ones (Kou et al., 2017). Knowing how many riders are entering or queuing in a station can justify more rail cars, increased train frequency or other ways to keep riders on time.

Another problem with overcrowding is reduced sanitation and health concerns. As the number of riders entering or leaving a station grows, so do the health and environmental conditions for that station. Air quality can quickly degrade with increased riders (Xu & Hao, 2017). Riders are likely to avoid subways for fear and concern of viral transmissions such as Covid-19 transmission (Tan & Ma, 2020) and use other forms, such as private car to get. BART authorities want to keep crowd sizes down, and station queues as short as possible to reduce transmissions and understanding the number of riders expected at a station at different times of the day can help ensure proper health and safety protocols are in used. It is very likely viral transmissions of Covid-19 will remain a serious concern for many riders and could adversely affect ridership moving forward. Keeping BART stations clean, well sanitized is likely to be a top priority for riders and authorities together.

While BART is a governmental project, it operates like any profit and loss business. BART must continue to look for new ways to increase revenues. Targeted ads are one such way. Knowing how many riders are entering and leaving a given station on given day and hour and, even more important, where those riders are coming from and departing to, provides an opportunity for local business to advertise to a very targeted group of commuters (Ketelaar et al., 2018). Local business could target ads that are highly sensitive to the estimated number of riders who will be visiting or passing by their place of business at any day or hour (Vukasović, 2020). For example, a coffee shop could present advertisements at the origin station, knowing the number of riders heading towards their shop, and continue ads on the commute to further encourage customers to stop by and purchase a cup of coffee.

Offering “at station” and “on train” advertisements can be an improvement over traditional billboard advertisements that often distract drivers (Belyusar et al., 2016). It is shown that consumers in a crowded subway or station are, surprisingly, twice as likely to respond to advertisement than in in non-crowded cases (Andrews et al., 2016). This can present a great opportunity for local businesses to target stations with large volumes of riders passing through.

Finally, and importantly, BART is an extremely environmentally positive transportation service. It is estimated that each commuter, in comparison to a private vehicle, the equivalent of 224 miles per gallon on average and up to 422 miles per gallon during peak times (*Sustainability*, n.d.). Many countries and cities are transitioning from increasing road-based commutes to rail based to improve quality of life and reduce emissions all at the same time (Ćetković et al., 2020). The more riders are aware of the positive impact they are having by using public transportation services, perhaps the more encouraged riders will feel to continue using such an innovative service and experience the joy and stress-free commutes of public transportation (Ercan et al., 2016).

# **Future**

One limitation of this study, is Covid-19 effects. The purpose of this study is to look at the pattens riders take each day, and if those patterns are repeatable. It is likely data would need to be separated between pre and post Covid-19 periods. As BART schedules were changed and train availability was drastically reduced during Covid19, it is very possible that route patterns remained unchanged. This would only strengthen this research if accurate. Currently, this is not a hypothesis test in this study.

If riders are shown in this study to commute is consistent routes each day, it would be very interesting to see how those patterns are similar or changed during Covid-19 periods. Did riders still use the same commutes, just drastically reduced in numbers? How were riders affected?

**Conclusion**

While many subway and public transportation studies focus on the number of riders entering and exiting stations, few look at the ridership route distributions patterns. The distribution patterns tell us where the proportion of riders are heading each morning. Armed with this information, local businesses, riders and BART itself can offer improved services and increase the quality of riders’ commute each day.

Riders can be offered advertisements for local business they are soon to pass by. Local businesses can market their products and services to an anticipated number of riders. Those targeted ads can start the moment the rider enters the station and even continue on the train through the duration of the ride given increased advertisement exposure. Other services, such as connecting information or bus schedule and arrival times can be flashed ahead of riders that target the bulk of riders expected to make the transfer.

By better serving BART commuters, the more riders may be encouraged to take advantage of public transportation, thus reducing emissions and minimizing road congestion. BART is an incredible resource that can and can continue to improve and evolve to meet the needs and wants of riders and local businesses.

# References

Andrews, M., Luo, X., Fang, Z., & Ghose, A. (2016). Mobile ad effectiveness: Hyper-contextual targeting with crowdedness. *Marketing Science*, *35*(2), 218–233.

Belyusar, D., Reimer, B., Mehler, B., & Coughlin, J. F. (2016). A field study on the effects of digital billboards on glance behavior during highway driving. *Accident; Analysis and Prevention*, *88*, 88–96.

Ćetković, J., Lakić, S., Bogdanović, P., Vujadinović, R., & Žarković, M. (2020). Assessing environmental benefits from investment in railway infrastructure. *Polish Journal of Environmental Studies*. https://doi.org/10.15244/pjoes/109849

Ercan, T., Onat, N. C., & Tatari, O. (2016). Investigating carbon footprint reduction potential of public transportation in United States: A system dynamics approach. *Journal of Cleaner Production*, *133*, 1260–1276.

Ketelaar, P. E., Bernritter, S. F., van Woudenberg, T. J., Rozendaal, E., Konig, R. P., Hühn, A. E., Van Gisbergen, M. S., & Janssen, L. (2018). “Opening” location-based mobile ads: How openness and location congruency of location-based ads weaken negative effects of intrusiveness on brand choice. *Journal of Business Research*, *91*, 277–285.

Kou, W., Chen, X., Yu, L., Qi, Y., & Wang, Y. (2017). Urban commuters’ valuation of travel time reliability based on stated preference survey: A case study of Beijing. *Transportation Research. Part A, Policy and Practice*, *95*, 372–380.

*Sustainability*. (n.d.). Bart.Gov. Retrieved April 18, 2021, from https://www.bart.gov/about/projects/cars/sustainability

Tan, L., & Ma, C. (2020). Choice behavior of commuters’ rail transit mode during the COVID-19 pandemic based on logistic model. *Journal of Traffic and Transportation Engineering (English Edition)*. https://doi.org/10.1016/j.jtte.2020.07.002

Vukasović, T. (2020). Understanding the consumers’ personal characteristics as the starting point for targeted marketing. *Management*, *15*(1), 29–41.

Wang, X., Rodríguez, D. A., Sarmiento, O. L., & Guaje, O. (2019). Commute patterns and depression: Evidence from eleven Latin American cities. *Journal of Transport & Health*, *14*(100607), 100607.

Xu, B., & Hao, J. (2017). Air quality inside subway metro indoor environment worldwide: A review. *Environment International*, *107*, 33–46.