Aggregation approaches for GEWISS – “Street front” aggregation using ALKIS Addresses

Dochev I., Seller H., Peters I.

Technical Urban Infrastructure Systems Group Prof. Irene Peters

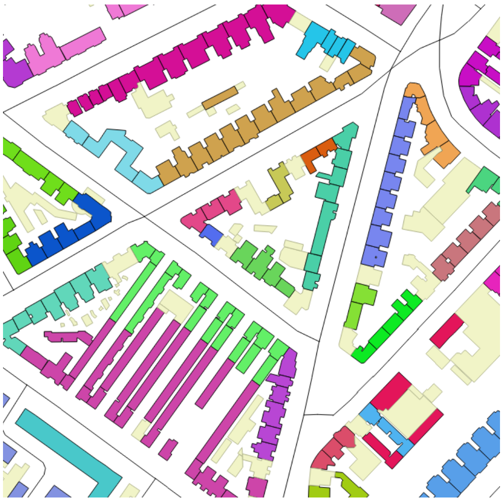
24.07.2017

Working paper
Please quote this working paper when using the methods or results herein

Summary

This paper presents a GIS-based method for grouping buildings from a digital cadaster in order to fulfill data protection requirements. It was prepared in the context of the Hamburg GEWISS and the Hamburg *Wärmekataster* projects*,* both of which include the construction of a thematic map depicting building heat demand (i. e., a heat demand cadaster). Since both projects provide for these maps to be made publicly available, general personal data protection requirements have to be met – in this case, no information that can be traced to an individual private household may be published. This is operationalized by the "5+" rule, stating that these households have to be aggregated into five or more units to preserve their anonymity.

In order to meet this requirement, we devised an aggregation level - "the Street Front" aggregation – which is finer than a *Baublock* and coarser than the individual building. It can be viewed as splitting the *Baublock* into parts, each overlooking a different street.



Example of "Street Front" aggregation. Black lines indicate streets, solid polygons indicate building footprints, colors depict aggregated building groups. Beige depicts non-residential buildings

The algorithm that creates this aggregation executes the following steps:

1. It filters out non-residential buildings
2. It assigns a minimum amount of dwelling units for each individual building, based on building characteristics in the digital cadaster (ALKIS)
3. Based on *Baublock* identifier (e. g., "101020" in the ALKIS) and street name of a building (e.g. "Gertrudenkirchof" in the ALKIS), it assigns a group identifier for each building (e.g. '*101020_Gertrudenkirchof*')
4. For every building group thus produced, if the amount of dwellings in the group is less than five, the group is merged with the nearest group in the same *Baublock* with more than five dwellings (by changing the group identifier to the one of the bigger group)
5. If there is no big group in the *Baublock,* but the sum of all dwellings in the *Baublock* is larger than five, then the entire *Baublock* constitutes one group. (e.g. '*101020')*
6. If there are less than five dwellings in the *Baublock*, the entire *Baublock* is anonymized and no data is published for it (the group identifier for the buildings is then 'BB_Anonymized'*)*

I. Introduction

The method described in this paper aggregates building information for the purpose of personal data protection, a principle which in Germany is written into law at the federal and at the *Land* levels. Personal data protection in general does not allow for publishing information that can be traced to an individual household. This requirement is operationalized by the "5 +" rule, stating that households have to be aggregated into five or more units to preserve their anonymity.[1]

The context in which this method was developed is the GEWISS project in Hamburg, Germany, one of whose tasks is the preparation of a thematic map depicting building heat demand. This thematic map is the basis for the Hamburg Heat Demand Cadastre (*Wärmekataster*) published by the Hamburg Ministry for Environment and Energy (*Hamburger Behörde für Energie und Umwelt, BUE*). Since this Heat Demand Cadaster is considered a planning and decision support tool for urban heat supply planning in conjunction with urban development, the aggregation method should serve these planning purposes, in addition to sufficing data protection requirements.

The underlying logic of the method is that if one views building heat demand maps from the perspective of the "heat supply planner", it makes sense to group buildings based on the street that they overlook, for the pipes of a heating grid often follow the street network. On the other hand, urban blocks (*Baublöcke*[2]) are an official spatial unit, defined in the Hamburg cadaster (ALKIS) for which also some statistical data is collected in the course of the Census. Therefore, it is preferable choosing an aggregation that is consistent with the *Baublock* level.

For this reason, all of the buildings which look onto the same street are grouped according to the *Baublock* to which they belong. This is equivalent to dividing the *Baublock* into groups, each of which overlooks one of the streets surrounding it. Figure 1 depicts an example.

This way of grouping buildings is referred to as "Street Front" grouping further in this paper. Since no group includes buildings from more than one *Baublock*, the defined "street front" groups can always be aggregated up to the *Baublock* level if needed.

[1] This rule has been agreed with the Office of the Hamburg Commissioner for Data Protection and Freedom of Information (*Hamburger Datenschutzbeauftragter*)

[2] *Baublock*" is an area bordered by city streets. The entire built-up area of a city is partitioned into contiguous *Baublöcke*.

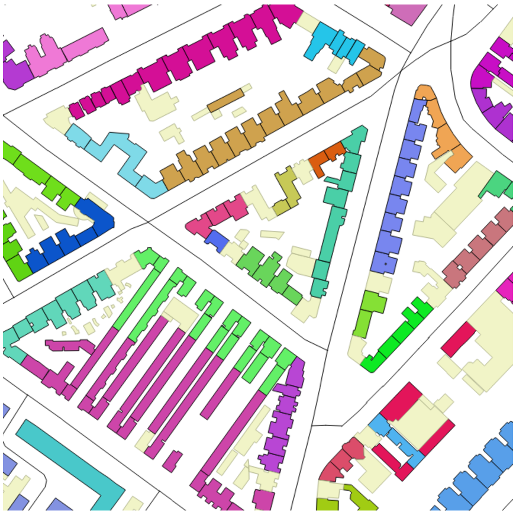


Figure 1. Example of "Street Front" aggregation. Black lines indicate streets, solid polygons indicate building ground floors, colors depict aggregated building groups. Beige depicts non-residential buildings.

The specific data protection requirements stipulate that no group should contain less than five dwelling units. In the usual case, the "Street Front" aggregation groups contain 5 to 20 buildings (this can also be seen in Figure 1), which is more than enough to satisfy data protection. However, there are exceptions that have to be dealt with. The logic adopted to do this is to merge "small" groups (containing less than five dwelling units) with "big" ones (containing more than five dwellings units). The aggregation method therefore comprises of the following steps:

1. Filter out non-residential buildings. Currently only residential buildings are considered.

2. Assign a minimum amount of dwelling units for each individual building, based on building characteristic in the digital cadastre (ALKIS).

3. Assign a group identifier for each building (e.g. '*101020_Gertrudenkirchof',* where '101020' is the *Baublock* identifier in the ALKIS and "Gertrudenkirchof" the name of the street).

4. For every group thus produced, if the sum of dwellings in the group is less than five, the group is merged with the nearest group in the same *Baublock* with more than five dwellings (the group identifier is changed to the one of the "big" group).

5. If there is no big group in the *Baublock,* but the sum of all dwellings in the *Baublock* is larger than 5, then the entire *Baublock* becomes one group (e.g. '*101020')*.

6. If there are less than five dwellings in the *Baublock*, the entire *Baublock* is anonymized and no data is to be published for it (building identifier = 'BB_Anonymized')

7. Additionally, if a group of buildings is "small", but too far away from a "big" group with which it could be merged, or there are large non-residential buildings in between, the group is also anonymized (building identifier = 'Anonymized')

What these steps omit is the attempt to divide "big" clusters into smaller "big" clusters wherever possible in order to achieve a finer aggregation (e.g. if a cluster has a minimum of ten dwelling units, then it can be divided into two groups of five). This was not implemented for this algorithm, but would constitute an improvement over the current version, and we are working in this direction.

II. Building Functions

As mentioned, this paper covers an aggregation method that currently deals with residential buildings only. However, the script that was prepared can take as input all buildings in the digital cadaster and filter them, using only the residential ones, and producing a blank string for all others. This was done in order to avoid the need for manual "pre-filtering" of buildings and thus making it easier for the user. The filtering is the first step that the script carries out, and it is based on an input of all building function codes that are considered residential. The digital cadaster of Hamburg (ALKIS[3]) designates a building function for (almost[4]) every building in Hamburg. These building functions are grouped in three general classes: residential (code 1xxx), commercial and industrial (2xxx) and public buildings (3xxx)[5]. Even though the general class 1000 is considered residential, some of the building functions therein were omitted for one of two reasons – the function implies very low or no heat demand; or the function implies a different type of residential that affects heat demand, e.g. a home for the elderly or a student's dorm. Table 1 (based on the GeoInfoDok, the official documentation of the ALKIS) summarizes the building functions currently considered ("default setting" for the algorithm). These functions can be changed at the beginning of the algorithm.

Wohngebäude 1000 (G)	Functions that are included in the heat demand calculation and subsequent aggregation. Mixed-use buildings are included and taken as „residential" due to the "dominance" principle in the ALKIS, stating that these are "predominantly" residential.
Wohnhaus 1010	
Gemischt genutztes Gebäude mit Wohnen 1100	
Wohngebäude mit Gemeinbedarf 1110	
Wohngebäude mit Handel und Dienstleistungen 1120	
Wohn- und Verwaltungsgebäude 1121	
Wohn- und Bürogebäude 1122	
Wohn- und Geschäftsgebäude 1123	
Wohngebäude mit Gewerbe und Industrie 1130	
Wohn- und Betriebsgebäude 1131	
Wohnheim 1020	Functions that are excluded from the heat demand calculation and subsequent aggregation. The main, predominant function of buildings with these function codes is considered not "strictly residential" from an energetic point of view and will be tackled when the other non-residential buildings are included.
Kinderheim 1021	
Seniorenheim 1022	
Schwesternwohnheim 1023	
Studenten-, Schülerwohnheim 1024	
Schullandheim 1025	
Land- und forstwirtschaftliches Wohn- und Betriebsgebäude 1220	
Bauernhaus 1221	
Wohn- und Wirtschaftsgebäude 1222	
Forsthaus 1223	
Gebäude zur Freizeitgestaltung 1310	
Ferienhaus 1311	
Wochenendhaus 1312	
Gartenhaus 1313	

Table 1. Overview of the residential building functions currently used.

[3] The ALKIS cadastral system is actually Germany-wide, but „ALKIS" in this paper refers to the Hamburg one.

[4] There is a very small number of buildings without a function entry in the cadastre.

[5] http://www.adv-online.de/AAA-Modell/Dokumente-der-GeoInfoDok/GeoInfoDok-7.0/

III. ALKIS Addresses

Lack of address

As previously described, the algorithm uses the street name of a building as found in the ALKIS in order to assign the building to a group. This has the disadvantage that not all buildings in the ALKIS have addresses, which however is not a serious issue. The residential buildings (filtered using the "default setting") are 217 892 (ALKIS 2016 Q1[6]). Although there are 22 574 (out of these 217 892) buildings that do not have an address, the average footprint area of these is 42 m^2, which is rather small and signals that these are smaller auxiliary buildings. Furthermore, the total gross floor area of these 22 574 buildings amounts to only 1.4% of the total gross floor area of all residential buildings (default setting). For this reason, omitting these buildings is not a serious issue.

Multiple addresses

Even though relatively straightforward, the relation between a building as defined in the ALKIS (basically a polygon) and the addresses associated with it is actually a "one-to-many" relation. This means that a building can have more than one address, depending on, in general, the number of entrances it has. For example, if a building is on "*Winsener Str*" and has three entrances – №18, №20, №22 or №18a, №18b, №18c or similar, this will mean three addresses associated with one building. These addresses may or may not have the same street name, for example when the building is a corner building and has one entrance on one street and another entrance on the other. For the current discussion this has two effects:

1. In the algorithm, the input has to be flattened, which currently means taking the street name of any one of the addresses associated with the building at random. This has no effect on buildings except the ones on street corners and even there whether a building is considered in one or the other group is not relevant for the purpose at hand.
2. The number of address points is considered a signal for the number of entrances which is a clue when estimating the minimum number of dwellings – a residential building of three entrances most likely has at least three dwellings. In reality this assumption might not hold true in each and every case, however the rules developed in the next chapter underestimate the amount of dwellings per building to a large extent (to be on the "safe" side) and can make up for any overestimation due to assuming that an address point equals an entrance. This assumption is also only made for building types considered to be multifamily buildings.

IV. Minimum Dwelling Units

Since the main reason for the need to group buildings is the data protection requirement of a minimum of five dwelling units per group, estimating this number for each building is crucial. There is no indication in the digital cadaster as to how many dwelling units are present in a given building, therefore assumptions have to be made according to other characteristics (e.g., number of entrances, number of floors, etc.).

[6] ALKIS 2016, first quarter of the year. A new version of the ALKIS is published on the Hamburg Transparency portal around every 3 months.

Since the requirement is that there should be a minimum of dwelling units and no upper bound, our approach for estimating this number for each building is to "underestimate" and be "on the safe side" rather than try to pinpoint with more precision and risk to overestimate (and then find out that we do not reach the "minimum of five dwelling units" criterion). Therefore, we estimate an utmost minimum, with the clear understanding that in most cases the true dwelling unit count will be a lot higher. An example of this conservative underestimation is that if a multifamily building is of "mixed-use" (so **not** use code 1010, which designates "pure" or "only" residential use) then it is assumed it has a minimum of one dwelling unit and no more. In most cases this is a gross underestimation, but since it is unclear how much of the building area is occupied by residential dwellings, any assumptions as to this could lead to overestimations, which could lead to groups of less than five dwelling units.

The basic assumption behind the dwelling count estimation is that it is highly unlikely that a dwelling spans over more than one floor of a building. The second basic assumption is that a dwelling most likely does not span across multiple entrances and therefore address points. As a precaution this is used only for "pure" residential multifamily buildings (code 1010 and *Bauweise*[7] codes: 1200 - *Freistehender Gebaeudeblock,* 2400 – *Gruppenhaus or* 2500 - *Gebaeudeblock in geschlossener Bauweise*). As a further precaution, although only "pure" residential function is considered to have more than one dwelling, the ground floor is also taken out, assuming it is possible that a mistake occurred in the ALKIS and a "pure" residential has non-residential uses on the ground floor. These assumptions and arguments translate into a classification algorithm summarized in Table 2.

Function	***Bauweise***	**Floors**	**Assumed minimum dwelling units**
any	any	0	0
Mixed-use	any	>0	1
"Pure" residential (code 1010)	1100/2100/2200/*	1 to 3	1
		3 to 5	3
		>5	(floors -1)
	1200/2400/2500	1 to 3	1 * Number of entrances
		3 to 5	3 * Number of entrances
		>5	(floors - 1)* Number of entrances
	unknown	1 to 3	1
	unknown	>3	(floors/2) * number of entrances

Table 2. Overview of minimum dwelling unit classification. *Single-family buildings (1100 *Freistehendes Einzelgebäude,* 2100 *Doppelhaushälfte* 2200 *Reihenhaus)* generally do not have more than 2 to 3 floors, however, the rules are more robust, so that if the *Bauweise* is wrong the algorithm knows how to handle exceptions.

V. Plot constraint

An additional option in the algorithm is to deny allocating buildings in the same plot into different groups. This was implemented in order to allow the visualization based on plot geometries and not building geometries.

[7] An ALKIS attribute

VI. Results

The result of the algorithm is a group identifier for each residential building (default setting). An overview of the size of the groups is presented in Figure 2. A visual example of the grouping in different urban settings is presented in Figure 3. The overall size of the groups is in the range between 5 to 20 buildings, although there are groups with significantly more. The visual inspection leads us to conclude that the results are plausible, however with slight "imperfections" in some places, mainly caused by corner buildings. These imperfections are more of aesthetical nature since the "5+" (of dwellings) is met for the whole dataset.

Non-residential buildings that are located in-between buildings in a single group also could be considered a problem for the visualization/usage of the map, however not as a result of the algorithm itself, but of the overall decision to not include them at this point. Modifying the default setting to include non-residential uses is a viable way to expand the results, but the percent of buildings lacking addresses in the ALKIS is larger for non-residential buildings, which limits the possibilities of this concrete approach.

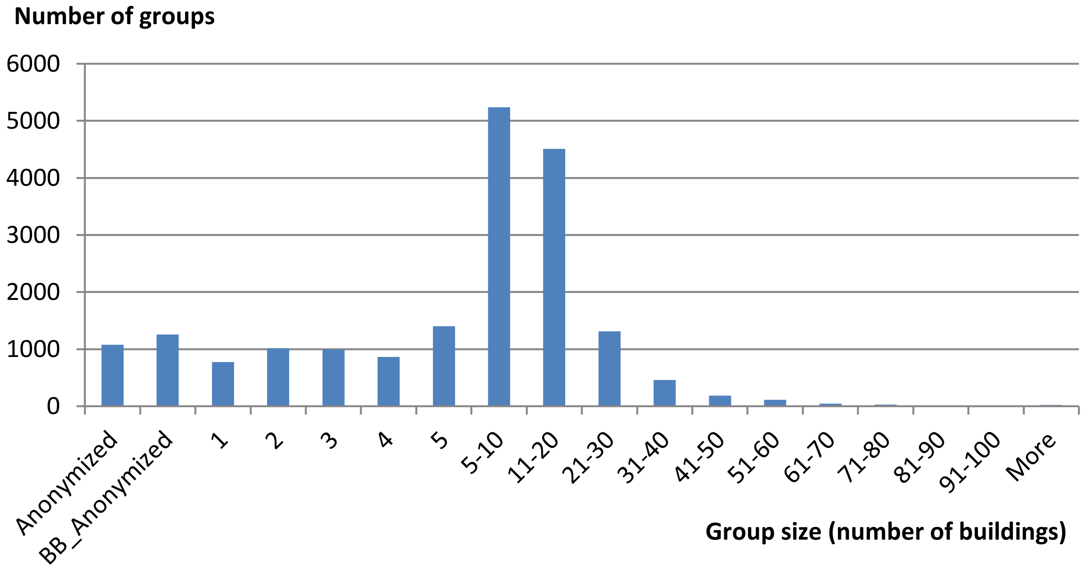


Figure 2. Frequency of building group sizes.

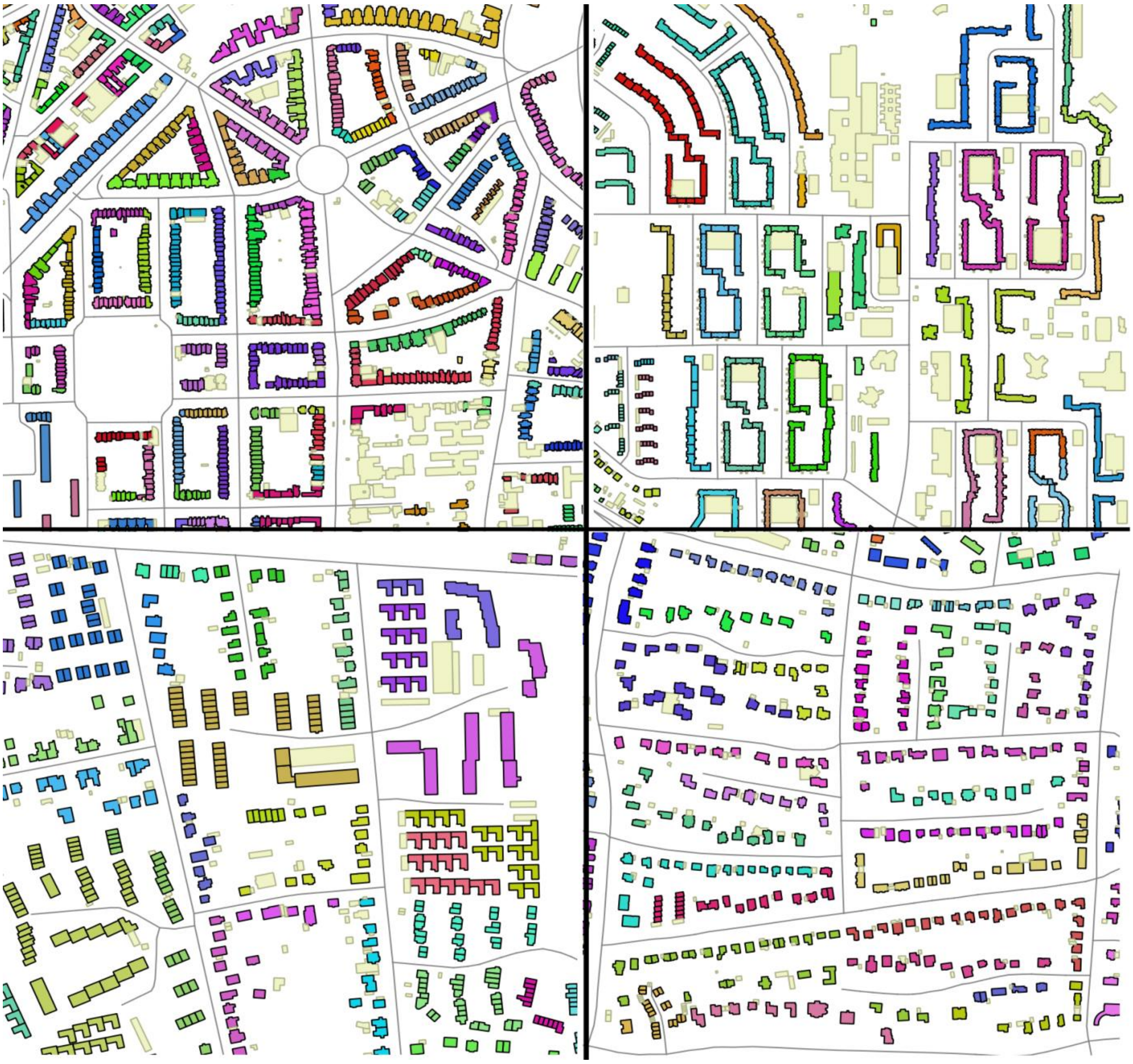


Figure 3. Examples of "Street Front" aggregation in different urban settings. Different colors depict the different groups. Beige depicts non-residential buildings. The examples are from *Harvestehude*, *Billstedt (Mümmelmansberg)*, *Harburg (Marmstorf)* and *Othmarschen* (clockwise starting from the upper left).

VII. Discussion and outreach

The presented method, although not perfect, is a plausible and viable way of ensuring that data protection requirements are met for thematic maps – in this particular case, thematic maps of heat demand.

The concrete visualization of the heat demand of the defined groups of buildings can be seen as a further task on its own. Nevertheless the aggregation algorithm provides for a relatively simple visualization based on plot geometries. By connecting buildings to their respective plots a contiguous geometry of neighbouring plots can be obtained which can then be visualized. The "plot constraint" (See Section Plot Constraint) allows for this by insuring that buildings can be grouped based on plots without the need to split plot geometries (when for example single plots contain buildings from multiple groups).

Of course, some problems may arise for non-contiguous groups of plots, for example when other buildings (which are not part of the thematic estimation and thus should not be part of forming aggregate groups) are located in-between the buildings of interest. This point constitutes possibly the biggest remaining challenge – integrating non-residential buildings and deciding how to approach them – this could be resolved by non-residential buildings being allowed to form groups with residential ones, or by making use of multi-layered maps, with different layers for different functions, or similar.

Currently, the first version of the *Wärmekataster* covers only residential buildings. The aggregation presented here was also designed only for these, but the integration of non-residential buildings is planned both in GEWISS and the *Wärmekataster* projects. The method as it is, however, may reach its limits at that point, due to the larger number of non-residential buildings without addresses, which are essential for the grouping. Of course, including non-residential buildings might also require different rules for data protection, depending on which data of non-private entities is to be protected.

An additional point to consider is the relative inflexibility when it comes to a desired (not minimal) amount of dwellings or buildings per group. Generally, refraining from splitting larger groups (of 10 to 20 buildings) into smaller groups can have a positive effect on the quality of the heat demand data. Since building-level estimations for large amounts of buildings relies on many assumptions, it can produce large discrepancies between estimates and reality for individual buildings. Grouping buildings together and presenting aggregated data can average out some of these necessary errors, thus increasing the accuracy at the aggregated level. Therefore there are benefits to having larger groups. However, an improvement would be to be able to set a desired amount of buildings or dwellings, so that the averaging-out effects could be better controlled.

Due to these shortcomings of the method presented here, we are currently (July 2017) exploring alternative approaches to the problem.